PATRÍCIA ISABEL FERREIRA MARQUES

AN EVOLUTIONARY PERSPECTIVE INTO THE ROLE OF KALLIKREINS (KLKs) IN MALE REPRODUCTIVE BIOLOGY

Tese de Candidatura ao grau de Doutor em Ciências Biomédicas submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Orientador – Doutora Susana Seixas

Categoria – Investigadora

Afiliação – Instituto de Investigação e Inovação em Saúde, Universidade do Porto (I3S); Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Ipatimup).

Coorientador – Doutor Victor Quesada

Categoria – Investigador

Afiliação – Departamento de Bioquímica y Biología Molecular de la Universidad de Oviedo

Coorientador – Maria de Fátima Gärtner

Categoria – Professora Catedrática

Afiliação – Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Research work coordinated by:

Financiamento:

Este trabalho foi financiado por Fundos FEDER através do Programa Operacional Factores de Competitividade – COMPETE e por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto FCOMP-01-0124-FEDER-028251 (Refª FCT: PTDC/BEX-GMG/0242/2012). Este trabalho foi ainda financiado pela FCT através da atribuição de uma bolsa individual de doutoramento (SFRH/BD/68940/2010).

Ao abrigo do art.º 8º do Decreto-Lei n.º 388/70, fazem parte integrante desta dissertação os seguintes manuscritos já publicados, aceites para publicação ou em preparação:

Marques PI, Bernardino R, Fernandes T. Nisc Comparative Sequencing Program, Green ED, Hurle B, Quesada V, Seixas S. 2012. Birth-and-Death of *KLK3* and *KLK2* in primates: evolution driven by reproductive biology. Genome Biol Evol. 4(12): 1331-8.

Marques PI, Fonseca F, Sousa T, Santos P, Camilo V, Ferreira Z, Quesada V, Seixas S. 2015. Adaptive Evolution Favoring *KLK4* Downregulation in East Asians. Mol Biol Evol. *Epub ahead of print* (DOI: 10.1093/molbev/msv199).

Marques PI, Fonseca F, Carvalho AS, Puente DA, Damião I, Almeida V, Barros N, Barros A, Carvalho F, Mathiesen R, Quesada V, Seixas S. Rare and common variants in *KLK* and *WFDC* gene families and their implications into male infertility phenotypes. *In preparation*.

Em cumprimento do disposto no referido Decreto-Lei, a candidata declara que participou na obtenção, análise e discussão dos resultados, bem como na elaboração das publicações, sob o nome Marques PI.

“It is not our differences that divide us.
It is our inability to recognize, accept,
and celebrate those differences.”

Audra Lorde

Acknowledgments
Agradecimentos

In all these years, many people were directly or indirectly involved in shaping up my academic career. It would have not been possible for me to thrive in my doctoral work without their precious support.

Gostaria de agradecer à minha orientadora, Doutora Susana Seixas, por todo o apoio, entusiasmo e dedicação, não só neste desafio, mas também ao longo de todos estes anos que trabalhamos juntas. Sem dúvida, contribuiu para o meu crescimento profissional e pessoal. Por isso, por todos estes "atribulados" mas gratificantes anos, um grande obrigado!

To my co-supervisor, Doctor Victor Quesada, for accepting me as his student and for believing in me and in my work. Thank you for all the scientific discussions and the patience while teaching me bioinformatics (I know sometimes I can be stubborn). I must also thank you for being an amazing host during my stays in Oviedo and for all of the rides from and to the airport. Gracias por todo!

À Professora Doutora Fátima Gärtner, por ter aceite ser minha co-orientadora, por toda a ajuda e disponibilidade.

À Fundação para a Ciência e a Tecnologia, agradeço a concessão da bolsa de doutoramento SFRH/BD/68940/2010 sem a qual a realização deste trabalho não teria sido possível.

Ao Professor Doutor Eduardo Rocha, ao Doutoramento em Ciências Biomédicas e ao Instituto de Ciências Biomédicas Abel Salazar por me terem dado a oportunidade de realizar o Doutoramento numa área que me fascina.

Ao Professor Doutor Sobrinho Simões agradeço a oportunidade de desenvolver o meu trabalho nas excelentes condições oferecidas pelo Ipatimup e de integrar nesta notável equipa de "Ipatimupianos".

To Professor Carlos López-Otín, for giving me the opportunity to join an incredible team that allowed me to grow up as a scientist. I'm also thankful for his enthusiasm in science that certainly has caught me!

Gostaria também de agradecer ao Centro de Genética da Reprodução Prof. Alberto Barros (Professor Alberto Barros, à Dra. Joaquina e ao Dr. Nuno), ao Centro de Estudos de Infertilidade e Esterilidade (Professor Vasco Almeida e à Dra. Isabel Damião) e ao Jardim Zoológico de Lisboa (Dra. Teresa Fernandes e Dr. Rui Bernardino), pelo fornecimento de amostras sem as quais este trabalho não teria sido possível.

I would like to acknowledge all of the co-authors of the publications presented in this thesis for their contribution.

Ao grupo Genetic Diversity pela forma que me acolheu quando integrei no grupo e por terem criado um bom ambiente científico.

Um especial obrigado à Ana Lima, Andreia B., Andreia S., Catarina, Filipa, Joana, Marisa, Natália, Sílvia, Sofia e Zélia pela partilha de experiências (dentro e fora do Ipas), pelos bons momentos de descontração e por me terem ajudado (pelo menos tentaram) a manter a sanidade metal! Sem vocês não teria sido possível.

To all members of the Lopez-Otin lab, especially Ceci, David and Yaiza, for welcome me and make me feel like home.

À Tatiana, por tantos anos de amizade e por estar sempre presente quando precisei!

Aos meus pais, à minha irmã e ao Fred, pelo carinho, apoio incondicional e compreensão em todos os momentos.

Table of Contents

Figures List

Chapter 1

General Introduction

Chapter 3

Papers

Paper I – Birth-and-Death of *KLK3* and *KLK2* in Primates: Evolution Driven by Reproductive Biology

Paper II – Adaptive Evolution Favoring *KLK4* Downregulation in East Asians

Paper III – Rare and common variants in *KLK* and *WFDC* gene families and their implications into semen hyperviscosity and other male infertility phenotypes

Chapter 4

Final Discussion

Appendices

Appendix A – Supplementary Material Paper I

Appendix B – Supplementary Material Paper II

Appendix C – Supplementary Material Paper III

Tables List

Chapter 1
General Introduction

Chapter 3
Papers

Paper I – Birth-and-Death of *KLK3* and *KLK2* in Primates: Evolution Driven by Reproductive Biology

Paper II – Adaptive Evolution Favoring *KLK4* Downregulation in East Asians

Paper III – Rare and common variants in *KLK* and *WFDC* gene families and their implications into semen hyperviscosity and other male infertility phenotypes

Appendices

Appendix A – Supplementary Material Paper I

Appendix B – Supplementary Material Paper II

Appendix C – Supplementary Material Paper III

Abbreviations

1000G - 1000 Genomes Project

3D - three-dimensional

A

AD - Atopic dermatitis

ADAM2 - A disintegrin and metalloprotease domain 2

AI2A1 - Hypomaturation-type amelogenesis imperfecta

AR - Androgen receptor

ASN - Asians (CHB+JPT)

ASW - African ancestry in southwest United States of America

B

bp - Basepair

BTB/POZ - Broad complex Tramtrack bric-a-brac/Pox virus and zinc finger

BWA - Burrows-Wheeler Alignment

C

CDH13 - Cadherin 13

cDNA - complementary deoxyribonucleic acid

CDSN - Corneodesmosin

CEU - Utah residentswith ancestry from northern and western Europe

CFTR - Cystic fibrosis transmembrane conductance regulator

CHB - Han chinese in Beijing, China

CHS - Southern Han chinese in China

cKLK - chimeric Kallikrein

CLM - Colombian in Medellin Colombia

CLR - Composite likelihood ratio

cM/Mb - Centimorgan per megabase

CMS - Composite of multiple signals

CNV - Copy number variation

CTCF - CCCTC-Binding Factor

D

DAF - Derived allele frequency

DAZL - Deleted in azoospermia-like

DNA - Deoxyribonucleic acid

DIND - Derived Intra-allelic Nucleotide Diversity

DMEM - Dulbecco's Modified Eagle Medium

DPP1 - Dipeptidyl peptidase I

DSC1 - Desmocollin 1

DSG1 - Desmoglein 1

E

ECL - Enhanced chemiluminescence

EDAR - Ectodysplasin A receptor

EHH - Extended haplotype homozygosity

ENAM - Enamelin

EPPIN - Epididymal protease inhibitor (also known as *SPINLW1* and *WDFC7*)

eQTL - expression Quantitative Trait Loci

F

f - allele frequency

FIN - Finnish in Finland

FN - Fibronectin

FOS - FBJ murine osteosarcoma viral oncogene homolog

FOSL2 - FOS-like antigen 2

FOXP2 - Forkhead box protein P2

G

GBA3 - Glucosidase, beta, acid 3

GBR - British in England and Scotland

GTEx - Genotype-Tissue Expression project

GWS - Genome-wide scans

H

Het – Heterozygous

HKA - Hudson-Kreitman-Aguadé

HUGO - Human Genome Organization

HV – Hyperviscosity

Hz - Homozygous

I

IBS - Iberian populations in Spain

iHS - Integrated haplotype score

J

JUND - Jun D proto-oncogene

JPT - Japanese in Tokyo, Japan

K

kb - Kilobase

KLK - Kallikrein

KLK1 - Kallikrein1 (also known as pancreatic/renal kallikrein or tissue kallikrein)

KLK10 - Kllikrein 10 (also known as normal epithelial cell-specific or protease serine-like1)

KLK11 - Kallikrein11 (also known as hippostasin or serine protease 20)

KLK12 - Kallikrein 12 (also known as KLK-L5)

KLK13 - Kallikrein 13 (also known as KLK-L4)

KLK14 - Kallikrein 14 (also known as KLK-L6)

KLK15 - Kallikrein 15 (also known as prostinogen)

KLK1E2 - Equus caballus glandular kallikrein precursor

KLK1P - Canis lupus familiaris kallikrein 1 pseudogene

KLK2 - Kallikrein 2 (also known as human glandular kallikrein-1 or tissue kallikrein-2)

KLK3 - Kallikrein 3 (also known as prostate-specific antigen)

KLK4 - Kallikrein 4 (also known as KLK-L1, enamel matrix serine protease 1 or prostase serine protease 17)

KLK5 - Kallikrein 5 (also known as KLK-L2 or stratum corneum tryptic enzyme)

KLK6 - Kallikrein 6 (also known as neurosine)

KLK7 - Kallikrein 7 (also known as stratum corneum chymotryptic enzyme)

KLK8 - Kallikrein 8 (also known as neuropsin)

KLK9 - Kallikrein 9 (also known as KLK-L3)

KLKP1 - Kallikrein pseudogene 1

KRT77 - Keratin 77, type II

L

LCT - lactase

LD - Linkage disequilibrium

LEKTI -Llymphoepithelial kazal type inhibitor

LG - Lamellar granules

LRH - Long-range haplotype

LWK - Luhya inWebuye in Kenya

M

MAF - Minor allele frequency

ME2 - Malic enzyme 2, NAD(+)-dependent, mitochondrial

ME3 - Malic enzyme 3, NADP(+)-dependent, mitochondrial

miRNA - micro ribonucleic acid

MK - McDonald-Kreitman

mL - Mililiter

MM - Multimale

MMP20 - Matrix metalloproteinase-20

mRNA - messenger ribonucleic acid

MS – mass spectrometry

MTRR - 5-methyltetrahydrofolate-homocysteine methyltransferase reductase

MXL - Mexican ancestry in Los Angeles, CA

mya - Million years ago

N

N/A – Not applicable

NGS – Next-generation sequencing

NS - Netherton syndrome

NV – Non-hyperviscosity

O

OCA2 - Oculocutaneous albinism II

P

PAR-2 - Protease-activated receptor-2

PCR - Polymerase chain reaction

PI3 - Peptidase inhibitor 3, skin-derived (also known as *WFDC14* or *ELAFIN*)

PLRP2 - Pancreatic lipase-related protein 2

PSA - Prostate-specific antigen (also known as KLK3)

PUR - Puerto Rican in Puerto Rico

R

r^2 – linear correlation coefficient

REHH - Relative extended haplotype homozygosity

RH - Recombination hotspot

RNA - Ribonucleic acid

RPTOR - Regulatory associated protein of MTOR, complex 1

RT-PCR - Reverse transcription polymerase chain reaction

S

SB - Stratum basale

SC - Stratum corneum

SDS-PAGE - Sodium dodecyl sulfate poly-acrylamide gel electrophoresis

SEMG - Semenogelins

SEMG1 - Semenogelin 1

SEMG2 - Semenogelin 2

SERPINA5 - Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5

SERPINB11 - Serpin peptidase inhibitor, clade B (ovalbumin), member 11

SERPINB3 - Serpin peptidase inhibitor, clade B (ovalbumin), member 3

SERPINB4 - Serpin peptidase inhibitor, clade B (ovalbumin), member 4

SFS - Site frequency spectrum

SG - Stratum corneum

SIGLEC5 - Sialic acid binding Ig-like lectin 5

SIGLEC6 - Sialic acid binding Ig-like lectin 6

SLC24A5 - Solute carrier family 24 (sodium/potassium/calcium exchanger), member 5

SLC45A2 - Solute carrier family 45, member 2

SLPI - Secretory leukocyte peptidase inhibitor (also known as *WFDC4*)

SMIPS - Somatic Mutation Identification in Pooled Samples

SNP - Single nucleotide polymorphism

SNV - Single nucleotide variant

SPINK5 - Serine protease inhibitor Kazal-type 5

SPINLW1 - Serine protease inhibitor-like with Kunitz and WAP domains 1 (also known as EPPIN)

SRY - Sex determining region Y

SS - stratum spinosum

T

TGM4 - Transglutaminase 4

TRH - thyrotropin-releasing hormone

TYRP1 – Tyrosinase-related protein 1

TSI - Toscani in Italia

U

UM - Unimale

UTR - Untranslated region

UV-light - Ultraviolet light

V

VEP - Variant Effect Predictor

W

WFDC - whey acidic protein four-disulfide core domain

WHO - World Health Organization

WNT10A - Wingless-type MMTV integration site family, member 10A

X

XP-CLR - Cross-population composite likelihood ratio

XP-EHH - Cross-population extended haplotype homozygosity

Y

YRI - Yoruba in Ibadan, Nigeria

Z

Zn^{2+} - Zinc

ZP - zona pellucida

ZP2 - Zona pellucida glycoprotein 2

ZP3 - Zona pellucida glycoprotein 3

Abstract

The advances in sequencing technologies have greatly contributed to the increasing number of available genomes for different species, as well as to the development of detailed catalogs of human genetic variation. Likewise, these represent important tools for detecting footprints of natural selection acting on different timescales and help providing a better understanding of the molecular basis of current patterns of human disease susceptibility. Among the myriad of targets of natural selection, genes involved in reproductive functions are particularly relevant since they are at the frontline of the individual fitness.

The *kallikrein* (*KLK*) gene family encodes 15 serine proteases, often co-expressed in a wide variety of tissues and in many biological fluids, including in the seminal plasma. In this sense, KLKs are known to modulate key physiological processes through complex proteolytic cascades, in which family members act at different hierarchical levels. In the particular case of the semen liquefaction cascade, KLKs are involved in the hydrolysis of the major seminal structural proteins, the semenogelins (SEMGs), resulting in sperm release in the vaginal cavity. Previous studies provided evidence of KLKs substrates being preferred targets of natural selection through mechanisms linked to male fertility and sperm competition. Also, the loss of *KLK2* in some primate species was found to correlate to different semen physiologies. In addition, aberrant expression of most human KLK members was reported in individuals with abnormal semen parameters.

The three independent studies presented in the present work were all based on the central hypothesis that *KLK* genes might have been targeted by natural selection and that their genetic variation may underlie both beneficial and disease phenotypes. The analysis of the selective pressures acting on the *KLK* cluster was performed in two stages. The first was based on comparative and phylogenetic approaches, centered in *KLK2* and *KLK3*, for a total of 22 primate species. The second enclosed a comprehensive evaluation of the 1000 Genomes phase I data, in order to characterize a potential signature of natural selection among *KLK* genes in Asian populations. At the interspecific level, this study supported the origin of *KLK3* in Catarrhini through an event of *KLK2* duplication and functional divergence of *KLK3* towards a different substrate specificity and it unraveled an intricate evolutionary dynamics of *KLK2* and *KLK3* correlated to semenogelin gene structure, primate mating system and semen coagulation rates. On the other hand, in human populations, a complex signature of recent positive selection in East Asians was disclosed and characterized by a high frequency haplotype defined by three variants (rs1654456_G, rs198968_T and rs17800874_A) acting synergistically to promote *KLK4*

downregulation, which may be connected to tooth and epidermal features typical of these populations rather than with reproductive functions.

In the last part of this work, the variation of *KLK* genes, their substrates (semenogelins) and potential inhibitors of the whey acidic protein four-disulfide core domain (*WFDC*) locus was assessed in the scope of male infertility. This analysis revealed a higher burden of functional low-frequency variants in cases than in controls, independently of the considered infertility phenotype, but solely for the *KLK* cluster. Furthermore, it was possible to identify a significantly increased risk of semen hyperviscosity and asthenozoospermia associated with the variants rs61742847 (*KLK12*, p.P34L) and rs147894843 (*SEMG1*, p.G400D), respectively. In addition, other 12 nucleotide variants were also overrepresented in cases but, due to the still limited number of samples used in the genotype surveying, those were not statistically significant. Conversely, a decreased risk of hyperviscosity and oligozoospermia was observed for rs1654526 in *KLK7* and for a copy number variation in *SEMG1*, correspondingly.

Altogether, this work further supports that genes involved in proteolysis and reproductive biology, such as *KLK* genes, were targets of natural selection in the short and the larger timescales of human and primate evolution, respectively, and that *KLK* genetic variation may also underlie deleterious variants possible contributing to a lower reproductive fitness in humans.

Resumo

Os progressos nas tecnologias de sequenciação têm sido um forte contributo para o aumento do número de genomas disponíveis para diferentes espécies, bem como para o desenvolvimento de catálogos detalhados da variação genética humana. Deste modo, estes representam importantes ferramentas para a deteção de evidências de seleção natural em diferentes escalas temporais e ajudam ainda a obter um maior conhecimento das bases moleculares dos atuais padrões de suscetibilidade a doenças humanas. Entre os vastos alvos de seleção natural, os genes envolvidos em funções reprodutivas são particularmente relevantes, uma vez que estão na vanguarda da aptidão de um indivíduo.

A família de genes das calicreínas (*KLK*) codifica 15 proteases de serina, frequentemente co-expressas numa grande variedade de tecidos e em muitos fluidos biológicos, incluindo no plasma seminal. Neste sentido, as KLKs são conhecidas por modular importantes processos fisiológicos através de cascatas proteolíticas complexas, em que membros desta família atuam em diferentes níveis hierárquicos. No caso particular da cascata de liquefação do sémen, as KLKs estão envolvidas na hidrólise das principais proteínas estruturais, as semenogelinas (SEMGs), resultando na libertação dos espermatozóides dentro da cavidade vaginal. Estudos anteriores mostraram que os substratos das KLKs foram alvos de seleção natural através de mecanismos relacionados com a fertilidade masculina e competição espérmica. Também a perda da *KLK2* em algumas espécies de primatas está correlacionada com diferentes fisiologias do sémen. Além disso, foi reportada uma expressão aberrante de KLKs humanas em indivíduos com parâmetros seminais anormais.

Os três estudos independentes apresentados no presente trabalho foram baseados na hipótese central de que genes das *KLKs* podem ter sido alvo de seleção natural e que a sua variação genética pode contribuir para fenótipos benéficos e de doença. A análise das pressões seletivas no agrupamento das *KLKs* foi realizada em duas etapas. A primeira baseou-se em abordagens comparativas e filogenéticas, centradas na *KLK2* e na *KLK3*, para um total de 22 espécies de primatas. A segunda compreendeu uma avaliação abrangente dos dados da fase I dos 1000 genomas, com a finalidade de caracterizar uma potencial assinatura de seleção natural entre genes das *KLKs* em populações asiáticas. A nível interespecífico, este estudo suporta a origem da *KLK3* nos Catarríneos através de um evento de duplicação da *KLK2* e divergência funcional da *KLK3* para uma diferente especificidade de substrato e desvenda uma intrincada dinâmica evolutiva da *KLK2* e da *KLK3* correlacionada com a estrutura génica das semenogelinas, sistemas de acasalamento nos primatas e rácios de coagulação do

sémen. Por outro lado, nas populações humanas, foi revelada uma complexa assinatura de seleção positiva recente nos asiáticos de leste, caracterizada por um haplótipo de elevada frequência, definido por três variantes (rs1654456_G, rs198968_T e rs17800874_A) que atuam sinergicamente para promover uma redução nos níveis de *KLK4*, o que pode estar relacionado com características dentárias e epidérmicas típicas dessas populações e não com funções reprodutivas.

Na última parte deste trabalho, foi avaliada no âmbito da infertilidade masculina a variação dos genes das *KLKs*, dos seus substratos (semenogelinas) e potenciais inibidores do *locus whey acidic protein four-disulfide core domain* (*WFDC*). Esta análise revelou uma maior carga de variantes funcionais de baixa frequência nos casos do que nos controlos, exclusivamente no agrupamento das *KLK* e independentemente do fenótipo de infertilidade considerado. Além disso, foi possível identificar um maior risco significativo de hiperviscosidade do sémen e de astenozoospermia associado com os variantes rs61742847 (*KLK12*, p.P34L) e rs147894843 (*SEMG1*, p.G400D), respetivamente. Adicionalmente, outros 12 variantes nucleotídicos encontravam-se também sobre-representados em casos mas, devido ao limitado número de amostras utilizadas na genótipagem, estes não atingiram significância estatística. Por outro lado, foi observada uma diminuição do risco de hiperviscosidade e de oligozoospermia para o variante rs1654526 na *KLK7* e para uma variação do número de cópias (CNV) na *SEMG1*, correspondentemente.

De um modo geral, este trabalho suporta que genes envolvidos na proteólise e na biologia reprodutiva, tais como os genes das *KLKs*, foram alvos de seleção natural ao longo da escala evolutiva dos humanos e dos primatas. Este estudo suporta ainda que a variação genética das *KLKs* pode também conter variantes deletérios que poderão contribuir para um menor *fitness* reprodutivo nos humanos.

Chapter 1

General Introduction

1. Genetic variation and natural selection

Curiosity is part of the human nature and, like a child in the "why" stage, we always seek more knowledge about ourselves and our origins: "Where do we come from?", "Why are we so different from one another?", "Why are some individuals predisposed to certain diseases while others are not?". Answering these questions has been a major challenge in the scientific community for several decades and these have been the main focus of a considerable number of studies of human genetic diversity. In the last decades, a few milestones marked the field of human genetics. First, the *Human Genome Project* generated a human reference sequence, which provided improved knowledge about the location and organization of genes but bestowed little information on the genetic variation at a population level (Lander et al. 2001; Venter et al. 2001). This issue was later addressed in other initiatives such as *The International HapMap Project,* which genotyped millions of common single nucleotide polymorphisms (SNP) and assessed their allele frequencies and linkage disequilibrium (LD) patterns in three major human groups: Africans, Europeans and Asians (International HapMap 2005; International HapMap et al. 2007). More recently, the *1000 Genomes Project* extended the characterization of the human genetic variation to a total of 26 human populations using high-throughput sequencing approaches, thus allowing not only the identification of common single nucleotide variants (SNVs), but also rare SNVs as well as small and large insertions and deletions (Genomes Project et al. 2010; Genomes Project et al. 2012; Genomes Project et al. 2015; Sudmant et al. 2015). Altogether, these projects have provided a detailed catalogue of the human genetic diversity, which to this date represents a fundamental tool to address how different genes and variants influence current human traits and susceptibility to disease.

Importantly, the complex patterns of genetic diversity in present-day humans are the endpoint result of random mutation and recombination, demographic history and other evolutionary processes acting in different timescales, which may span from millions or thousands of years ago to more recent events occurred only a few generations ago. The patterns of human genetic variation, together with some archeological evidence, support the "out-of-Africa" model, proposing the origin of modern humans in East Africa followed by a series of expansions, bottlenecks and migrations between populations. The recent increment in genome-wide datasets of human genetic variation and the current sequencing efforts of archaic hominins have allowed researchers to confirm and refine this model. Moreover, these efforts have uncovered additional levels of complexity, such

as those associated to the dispersal of modern humans and their admixture with Neanderthals and Denisovans (Quintana-Murci et al. 1999; Voight et al. 2005; Gutenkunst et al. 2009; Green et al. 2010; Laval et al. 2010; Reich et al. 2010; Gravel et al. 2011; Harris and Nielsen 2013).

However, as humans spread out of Africa to other continents, they have been exposed to a novel set of environments, which included different climate variables (temperature, humidity, precipitation and short wave radiation flux), dietary components (roots, cereals and milk) and pathogen burden, consequently, humans had to adapt and survive under different selective pressures (Coop et al. 2009; Barreiro and Quintana-Murci 2010; Hancock et al. 2010b; Luca et al. 2010; Cagliani and Sironi 2013; Fumagalli and Sironi 2014; Karlsson et al. 2014). Natural selection is the evolutionary process by which heritable beneficial traits become more frequent in a population over time, as an outcome of a higher reproductive fitness of their carriers (Darwin 1859; Luca et al. 2010). Hence, a genetic variant associated with an advantageous phenotype will increase in frequency and it may eventually reach fixation by positive selection, whereas a deleterious variant will most likely be eliminated by purifying selection (Figure 1A) (Nielsen 2005; Sabeti et al. 2006; Luca et al. 2010; Scheinfeldt and Tishkoff 2013). Conversely, the balancing selection scenario favors diversity by maintaining two or more beneficial alleles in the population, thus none of the selected variants might reach fixation (Figure 1B) (Charlesworth 2006; Andres et al. 2009; Fumagalli et al. 2009; Andres et al. 2010; Ferreira et al. 2011; Ferreira et al. 2013b; Key et al. 2014; Teixeira et al. 2015).

Most approaches used to detect positive selection are based on distinctive signatures affecting the patterns of genetic variation, which may be identified individually by the analysis of the site frequency spectrum (SFS), patterns of LD decay, population differentiation or even using a combination of these metrics in the so-called composite methods (Table 1).

In a classical scenario of positive selection (or hard-sweep), the advantageous variant arises *de novo* in a population and is quickly swept to high frequencies, leaving in the genome a footprint of low diversity around the selected locus that fades over with genetic distance (Figure 1C). As this phenomenon occurs soon after the arising of the new variant, recombination will not have time to break down the haplotypic structure of the beneficial allele, generating unusual LD patterns and homogenous long-ranged haplotypes. Subsequently, many linked neutral sites will also be found at high frequencies as a result of *hitchhiking*. During the selective sweep, novel mutations may accumulate within the selected chromosomes, however, due to the rapid increase in frequency of

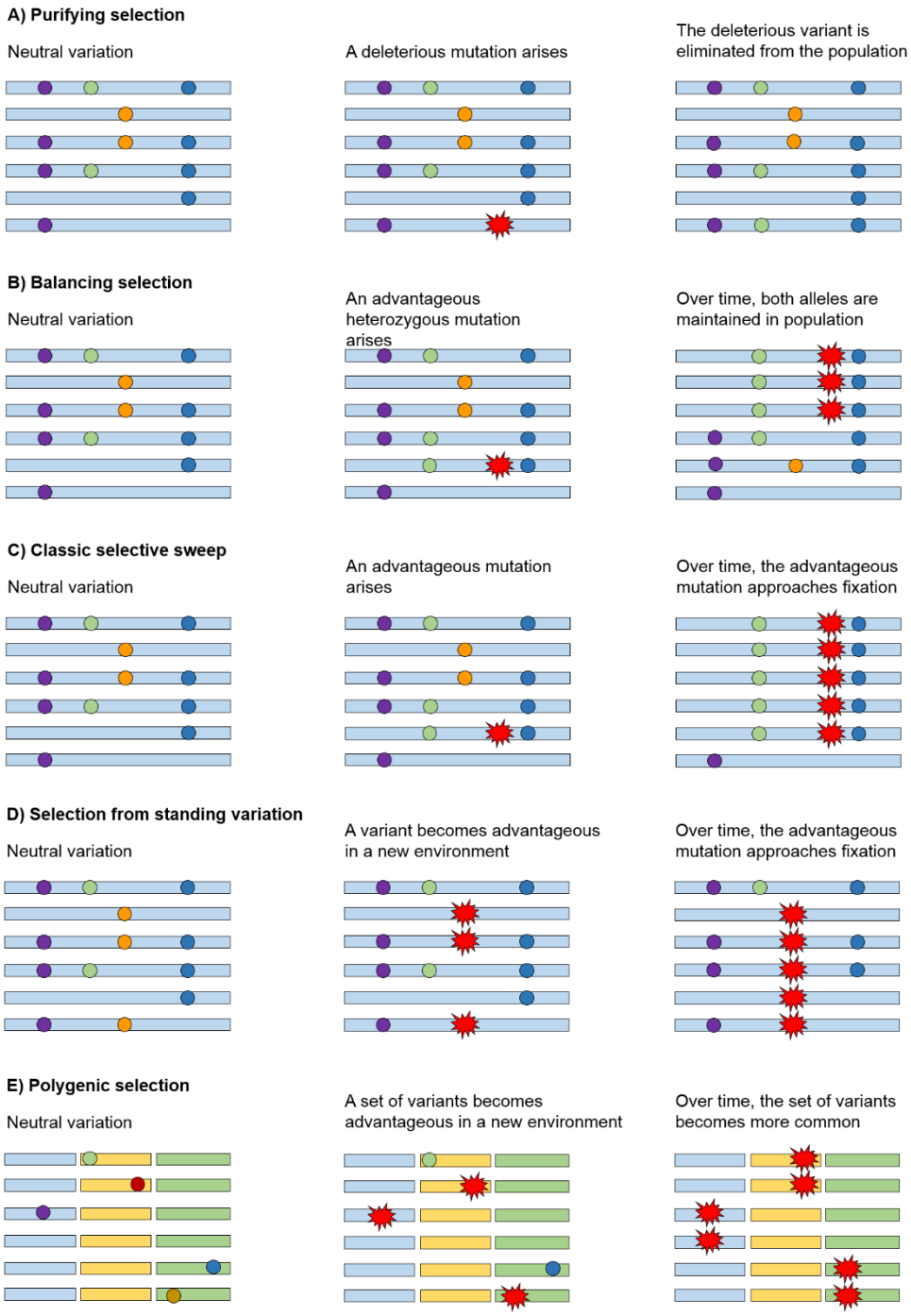


Figure 1 – **Different scenarios of natural selection.** Each panel depicts changes in variant frequencies over time. Horizontal blocks represent chromosomes, neutral variants are shown as circles on the chromosomes and advantageous or deleterious variants are represented with a 12-point star. **(A)** Purifying selection - in which deleterious variants are removed from the population. **(B)** Balancing selection - in which the two alleles are maintained in the population as a result of heterozygote advantage over homozygous individuals. **(C)** Classic selective sweep - in which a novel advantageous variant arises in a population and increases in frequency over time until it approaches fixation. **(D)** Selection from standing variation - in which a variant that is already present in the population becomes advantageous in a new environment and increases in frequency over time until it approaches fixation. **(E)** Polygenic selection - involves multiple loci in different chromosomes (represented by different colors), when a complex trait becomes advantageous, it increases in frequency as do the set of variants contributing to it (adapted from Scheinfeldt and Tishkoff 2013).

these chromosomes, most mutations will tend to be rare, causing an excess of rare variants (Luca et al. 2010; Scheinfeldt and Tishkoff 2013; Vitti et al. 2013). Furthermore, a selective advantage allele may be dependent on the specific environment in which it segregates and, since separated populations are likely to be subject to distinct selective pressures and local adaptation to different environments and cultural lifestyles, a particular allele may be beneficial in one population but not in others, generating frequency differences across populations (Excoffier 2002; Vitti et al. 2013). Examples of classical selective sweeps are related to lactose tolerance (*LCT*), skin pigmentation (*SLC24A5*) and tooth and hair morphology (*EDAR*) (Bersaglieri et al. 2004; Coelho et al. 2005; Lamason et al. 2005; McEvoy et al. 2006; Verrelli et al. 2006; Sabeti et al. 2007; Fujimoto et al. 2008; Basu Mallick et al. 2013).

Recent lines of evidence suggest that classic selective sweeps may have been rare in human evolution, whereas other mechanisms of selection are more likely to have been shaping current patterns of genetic diversity (Pritchard and Di Rienzo 2010; Pritchard et al. 2010; Hernandez et al. 2011; Crisci and Jensen 2012; Granka et al. 2012). One of such models proposes that selection may act on pre-existing variation that only becomes advantageous with sudden shifts in the environment (selection on standing variation or soft sweeps). In this scenario, the variant may already be linked to different haplotype backgrounds (Figure 1D) and, thus, the signature of selection may be more difficult to detect using conventional approaches designed to identify classic selective sweeps (Hermisson and Pennings 2005; Przeworski et al. 2005; Pennings and Hermisson 2006b; Pennings and Hermisson 2006a; Peter et al. 2012; Seixas et al. 2012; Turchin et al. 2012; Messer and Petrov 2013). In addition, such signals may be even harder to perceive when an adaptive trait is influenced by multiple loci, each one contributing with small effects (Figure 1E) (Pritchard et al. 2010). Regardless, some examples of polygenic adaptation (stature and pathogen immunity) and selection on standing variation associated with immunity (*SERPINB11*), climate (*ME3*, *ME2* and *RPTOR*) and dietary (*MTRR*, *PLRP2*, *GBA3*) adaptations are already described in the literature (Hancock et al. 2008; Hancock et al. 2010b; Lango Allen et al. 2010; Sun et al. 2010; Seixas et al. 2012; Daub et al. 2013).

Additional insights into the action of natural selection can be obtained by comparative genomic approaches, specifically in the case of ancient selective events, such as those underlying human speciation. Over a long timescale, positive selection can increase the fixation rate of beneficial alleles, which can be detected by comparing the rate of nonsynonymous substitutions (d_N) with the rate of synonymous substitutions (d_S) (Table 1). If there is no positive selection, synonymous and nonsynonymous changes

Table 1 - Common approaches used to detect selection (adapted from Vitti et al. 2013).

Approach	Intuition	Statistical tests
Comparative genomic methods	Synonymous substitutions are assumed to be selectively neutral. Thus, they can be used to assess the background rate of evolution. If the rate of nonsynonymous substitutions differs significantly, it is suggestive of selection.	d_N/d_S ratio (Yang 2007)
	Levels of polymorphism and divergence should be correlated (because both are primarily functions of the mutation rate) unless selection causes one to exceed the other.	McDonald-Kreitman (MK) (McDonald and Kreitman 1991) Hudson-Kreitman-Aguadé (HKA) (Hudson et al. 1987; Wright and Charlesworth 2004)
Frequency-based methods	In a selective sweep, a genetic variant reaches high prevalence together with nearby linked variants (high-frequency derived alleles). From this homogeneous background, new alleles arise but are initially at low frequency (excess of rare variants).	Tajima's *D* (Tajima 1989) Fu and Li *D** (Fu 1997) Fay and Wu's H (Fay and Wu 2000; Zeng et al. 2006)
Linkage disequilibrium-based methods	Selective sweeps bring a genetic region to high frequency in a population, including the causal variant and its neighbors. The associations between these alleles define a haplotype, which persists in the population until recombination breaks these associations down.	Long-range haplotype (LRH) (Sabeti et al. 2002) Integrated haplotype score (iHS) (Voight et al. 2006) Cross-population extended haplotype homozygosity (XP-EHH) (Sabeti et al. 2007)
Population differentiation-based methods	Selection acting on an allele in one population but not in another creates a marked difference in the frequency of that allele between the two populations. This effect of differentiation stands out against the differentiation with respect to neutral alleles.	F_{ST} (Excoffier et al. 2009)
Composite methods	Combining test scores for multiple sites across a contiguous region can reduce the rate of false positives.	Composite likelihood ratio (CLR) (Williamson et al. 2007) Cross-population composite likelihood ratio (XP-CLR) (Chen et al. 2010)
	Combining multiple independent tests at one site can improve resolution and distinguish causal variants. Different tests can provide complementary information.	Composite of multiple signals (CMS) (Grossman et al. 2010)

should occur at the same rate ($d_N/d_S = 1$), while under selection, might be an excess of nonsynonymous substitutions relative to the number of synonymous changes ($d_N/d_S > 1$). Other sophisticated methods may take into account the variation of d_N/d_S ratio among codons and evolutionary lineages, allowing to infer if specific sites have been targeted by positive selection or if selective pressures are acting on particular groups on the phylogeny or both (Nielsen and Yang 1998; Yang 1998; Yang et al. 2000; Yang and

Nielsen 2002; Bielawski and Yang 2003; Yang et al. 2005). These tools have been used in a wide variety of cases, for example, to reveal the presence of positive selection on genes involved in language and speech (*FOXP2*), reproduction (*ZP2*, *ZP3* and *ADAM2*) and immunity (*SIGLEC5*, *SIGLEC6*, *SERPINB3* and *SERPINB4*) (Swanson et al. 2001; Enard et al. 2002; Zhang et al. 2002; Kosiol et al. 2008; Gomes et al. 2014).

Recently, there is a growing body of evidence that regions of the human genome targeted by positive selection may also be associated with human disease. One possible explanation for this correlation is a change in the selective forces acting on human populations; alleles which were once beneficial may now be deleterious because the present environmental conditions are not the same as in the past ("thrifty genotype" hypothesis) (Neel 1962; Neel et al. 1998). This principle is well illustrated by variants predisposing to type-II diabetes. Several generations ago, fat accumulation was advantageous because it would largely raise the chances of survival in times of famine and, therefore, variants associated with this trait were positively selected. However, in modern societies, the current settled lifestyle and the ample food availability no longer render this characteristic beneficial and variants that were once valuable become maladaptive by increasing the risk of their carrier developing type-II diabetes (Di Rienzo and Hudson 2005; Di Rienzo 2006; Helgason et al. 2007; Crespi 2010; Vasseur and Quintana-Murci 2013). On the other hand, selective events could end in an antagonist pleiotropy, a phenomenon in which selection results in adaptation in one trait, or early in life, and in deleterious effects in other contexts or later in the lifespan (Clark and Swanson 2005; Nielsen et al. 2005; Corbo et al. 2008; Crespi 2010; Vasseur and Quintana-Murci 2013). Furthermore, when an advantageous variant rises in frequency during the selective sweep, the hitchhiking effect can drive disease-causing alleles to high frequency as well (Shiina et al. 2006; Huff et al. 2012). Therefore, there may be a close link between selection and disease. Thus, the combination of comparative and population genetic tools with functional information can be useful to exploit such disease candidate loci.

Overall, comparative genomics, human genome-wide scans (GWS) and candidate gene approaches have identified many potential examples of selection targets and verified that certain gene ontology categories, including sensory perception, dietary changes, immunity and host-pathogen interactions, reproduction and proteolysis, were enriched in genes under selection (Bustamante et al. 2005; Sabeti et al. 2006; Kosiol et al. 2008; Akey 2009; Bustamante and Ramachandran 2009; Grossman et al. 2013). In this context, the *kallikrein* (*KLK*) gene family represents a remarkable case for the study of evolution of proteolytic genes with implications in different biological processes including in reproductive biology and in human health and disease.

2. The *Kallikrein* (*KLK*) locus

The first kallikrein was identified in the 1930s as an abundant protease in the pancreas and, therefore, was named tissue kallikrein, based on the Greek word for pancreas “kallikreas”. Later work from independent groups led to the identification and characterization of 14 additional genes that now comprise the *KLK* gene family (Lilja 1985; Riegman et al. 1992; Gan et al. 2000; Yousef et al. 2000; Clements et al. 2001). As the members *KLK1, KLK2* and *KLK3* were the first ones described, they became known as “classical kallikreins” and *KLK1* remained as the prototypical *kallikrein* gene. In order to avoid confusion and harmonize the terminology of these genes, an official nomenclature system was proposed by the Human Genome Organization (HUGO). According to this nomenclature, *KLK1* is called *kallikrein-1*, whereas every other member of the family is termed *kallikrein*-related peptidase (Lundwall et al. 2006a). However, all these molecules are often simply named as *kallikreins*. Since their discovery, it became evident that *KLKs* are involved in many physiological processes and in several pathophysiologic conditions, as described below.

2.1. Structure and organization

The human *KLK* cluster, located at chromosome 19q13.3-13.4, spans over 265 kb and includes 15 paralogue genes coding for trypsin- or chymotrypsin-like serine proteases (*KLK1* to *KLK15*) and a transcribed pseudogene (*KLKP1*) (Yousef et al. 2004; Lu et al. 2006; Kaushal et al. 2008; Lundwall and Brattsand 2008; Prassas et al. 2015). The intergenic spacing between *KLK* genes is variable and ranges from approximately 1.5 to 32.5 kb, in which the smaller intergenic region is located between *KLK1* and *KLK15* and the largest one between *KLK4* and *KLK5*. Typically, genes extend from 4.4 to 10.5 kb, depending mainly on intron sizing, and, except for *KLK2* and *KLK3*, all of them are transcribed in the reverse chromosome strand. Moreover, all *KLK* genes display a common organization in five coding exons with variable 5’ and 3’ untranslated region (UTR) structures (Figure 2A) (Obiezu and Diamandis 2005; Lawrence et al. 2010). Furthermore, all *KLK* genes have at least one additional transcript resulting from alternative splicing, which in most instances are non-coding RNAs or result in shorter proteins with no protease activity (Kurlender et al. 2005; Koumandou and Scorilas 2013). Nonetheless, the function of these transcripts remains poorly understood and it has been suggested that they may mediate the signaling either at mRNA or at protein level (Koumandou and Scorilas 2013). The only exception to these structural features is *KLKP1*

which, in spite of having five transcribed exons, the *in silico* translation of the four mRNA transcripts indicate that none of them encode a serine protease and the two largest putative proteins are generated from a single exon (Yousef et al. 2004; Lu et al. 2006; Kaushal et al. 2008).

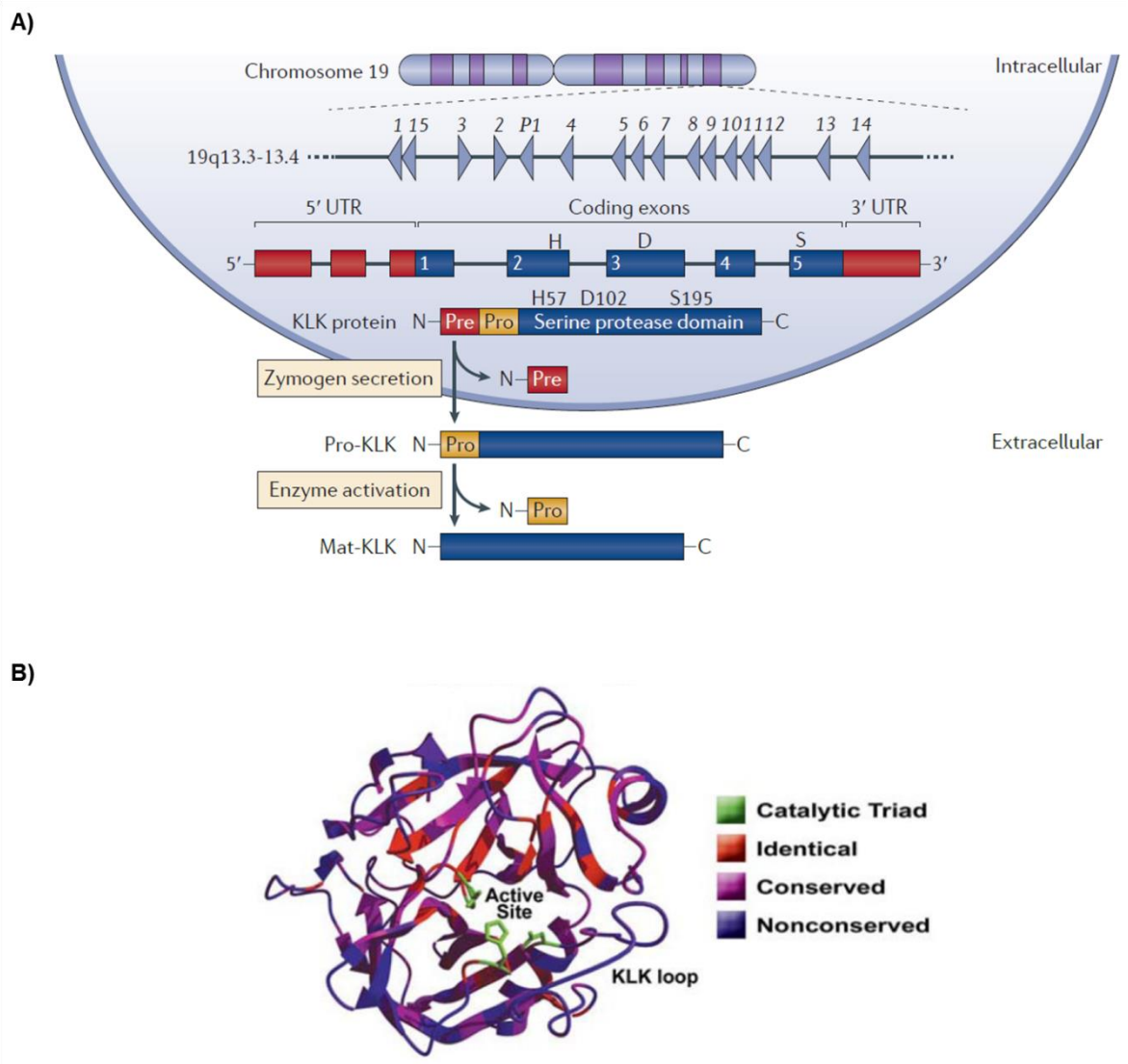


Figure 2 – Genomic and proteomic structure of KLK proteases. (A) The human *KLK* gene cluster is located at chromosome 19q13.3-13.4. Arrows show the relative position and the transcription orientation for the 15 coding genes and expressed pseudogene. In the mRNA scheme, the boxes and lines represent exons and introns, respectively. KLK proteins are expressed as pre-pro-enzymes, in which the pre-domain is required for intracellular trafficking and the pro-domain must be cleaved in order to generate a mature KLK. **(B)** The structure of a mature KLK based on the crystal structure of KLK1. The catalytic triad residues are shown in green. The position of the kallikrein loop is also shown. The amino acids that are identical among all kallikreins are in red, whereas those that are conserved in at least eight kallikreins are in purple. Non-conserved amino acids are in blue (adapted from Lawrence, Lai, and Clements 2010 and Prassas et al. 2015).

Kallikreins are encoded as single-chain pre-pro-enzymes (Figure 2A) of 244 to 293 residues long, sharing about 40-80% of protein identity, as well as a conserved catalytic triad of histidine (H57), aspartic acid (D102) and serine (S195) residues (standard chymotrypsin numbering) (Lundwall and Brattsand 2008). Besides the catalytic triad, amino acids involved in protein folding are also highly conserved among KLKs, whereas residues associated with substrate specificity are generally more divergent (Figure 2B). Once synthesized KLKs are directed to the endoplasmic reticulum by the signal peptide (the pre-domain), which is composed by 16 to 33 N-terminal amino acids, this pre-peptide is later cleaved in the secretory pathway yielding an enzymatically inactive pro-KLK. These molecules (zymogens) only become active upon the proteolytical removal of the pro-domain, which allows a conformational rearrangement of the protein three-dimensional structure by opening the KLK catalytic fissure. In most cases, the KLK pro-domain is cleaved after an arginine or lysine residue, indicating that they are activated by proteases with trypsin-like specificity including other KLKs, or in some instances by themselves (auto-activation) (Vaisanen et al. 1999; Brattsand et al. 2005; Michael et al. 2005; Memari et al. 2007; Yoon et al. 2007; Yoon et al. 2009; Lawrence et al. 2010). The exception to this rule is KLK4, which is cleaved after a glutamine residue and instead is activated by matrix metalloproteinase-20 (MMP20) and dipeptidyl peptidase 1 (DPP1) (Ryu et al. 2002; Tye et al. 2009; Yamakoshi et al. 2013).

So far, six human KLK structures have been solved (KLK1 and KLK3 to KLK7) and, according to those crystallographic models, KLKs are folded into two hydrophobic interacting sub-domains, each one comprising a six-stranded β-barrel and a α-helix, with the catalytic triad located at the interface between the two sub-domains (Figure 2B) (Bernett et al. 2002; Gomis-Ruth et al. 2002; Laxmikanthan et al. 2005; Debela et al. 2006; Debela et al. 2007a; Debela et al. 2007b; Debela et al. 2008; Menez et al. 2008). KLK substrate specificity depends on the residue that lies at the base of the substrate binding pocket, and it is further refined by the composition of the eight loops that surround the active site (Debela et al. 2008; Lawrence et al. 2010). In addition, KLK1, KLK2 and KLK3 structures also contain an 11-amino acid insertion known as the “kallikrein loop”, which is believed to play an important role in the substrate and inhibitor specificity (Yousef and Diamandis 2001; Borgono and Diamandis 2004; Laxmikanthan et al. 2005; Lundwall and Brattsand 2008).

2.2. Phylogenetic evolution

The structure of *KLK* genes and their organization within a single syntenic locus suggest that all members of this family have evolved from a common ancestor by a series of gene duplication events that occurred at different moments of vertebrate evolution (Figure 3). First, it was suggested that the origin of the *KLK* family arose before the marsupial-placental split, approximately 125-175 million years ago (mya) (Elliott et al. 2006). However, the growing number of genomes available in public databases have enabled a better resolution of the *KLK* cluster evolutionary history, tracing its origins to a tetrapod lineage approximately 330 mya (Pavlopoulou et al. 2010; Koumandou and Scorilas 2013; Kawasaki et al. 2014). Moreover, the identification of 11 marsupial orthologs (*KLK5* to *KLK15*) places the majority of duplication events prior to the separation of marsupial (Metatheria) and placental mammals (Eutheria) (Pavlopoulou et al. 2010; Koumandou and Scorilas 2013; Kawasaki et al. 2014). Consistently, several *KLK* orthologs were described for the platypus genome (*Ornithorhynchus anatinus*), but their tandem organization in a single syntenic cluster could not be assessed due to the lack of contiguous genomic segments and a still incomplete genome assembly (Pavlopoulou et al. 2010; Koumandou and Scorilas 2013; Lundwall 2013) (Figure 3).

Although the phylogenetic relationships among *KLK5* to *KLK15* genes are not yet well resolved, most studies done so far seem to agree that a single duplication encompassing *KLK9* and *KLK10* yielded *KLK11* and *KLK12*, or vice-versa (Olsson et al. 2004; Elliott et al. 2006; Lundwall et al. 2006b; Pavlopoulou et al. 2010; Lundwall 2013). Conversely, the events that generated *KLK2*, *KLK3* and *KLK4* are thought to have happened only after the placental mammals split (Elliott et al. 2006; Pavlopoulou et al. 2010; Koumandou and Scorilas 2013; Lundwall 2013). Initially, the absence of *KLK4* in elephant, hyrax, tenrec and armadillo genomes, together with phylogenetic data, suggested that this gene had arisen by a duplication of *KLK5* in the Boreoeutheria lineage (Elliott et al. 2006; Pavlopoulou et al. 2010; Koumandou and Scorilas 2013; Lundwall 2013). Still, the recent identification of a fragmented *KLK4* pseudogene in the genome of cape golden mole (*Chrysochloris asiatica*; Afrotheria) points to an earlier duplication event that most likely occurred in a common ancestor of placental mammals (Kawasaki et al. 2014). Further comparative analyses show even more complex histories for the classical *KLKs*, which display several events of gene gain and loss. For instance, both the dog (*Canis lupus familiaris*) and horse (*Equus caballus*) genomes present an additional *KLK1* homolog but, while in the horse this gene remained functional (*KLK1E2*), in the dog it degenerated into a pseudogene (*KLK1P*). On the other hand, in rodents, a remarkable

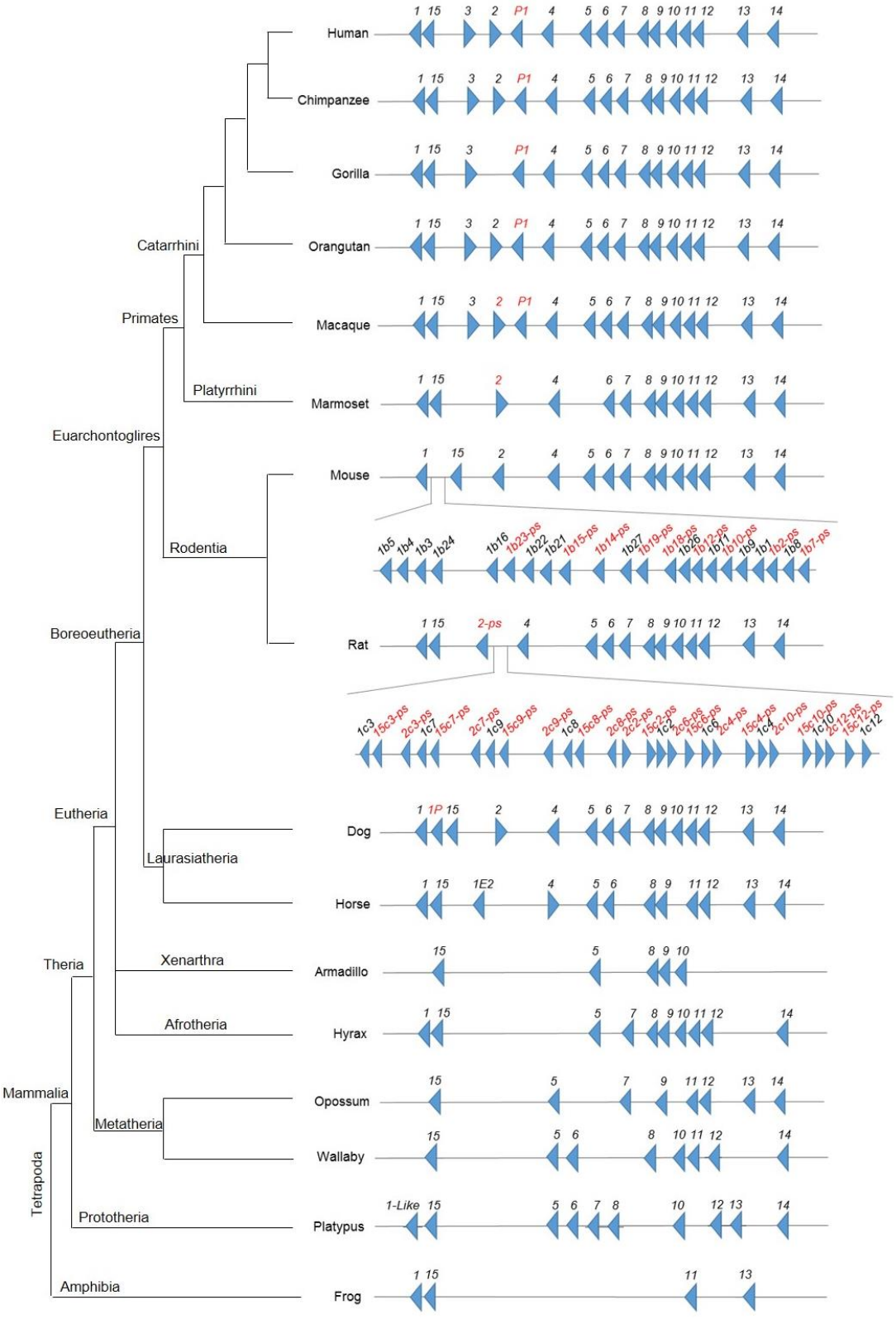


Figure 3 – Schematic representation of *KLK* genes in different species. The arrows specify the direction of transcription and known pseudogenes are indicated in red. Loci are not drawn to scale and bars do not represent chromosomes, as for many species the genomes are not yet fully assembled. On the left, a NCBI taxonomy-based dendrogram shows the taxonomic classes and the evolutionary relationship among taxa. Data compiled from Pavlopoulou et al. 2010, Koumandou and Scorilas 2013 and Lundwall 2013.

number of gene duplications generated an expanded locus with 13 and 10 *Klk1* paralogs in mouse (*Mus musculus*) and rat (*Rattus norvegicus*), respectively, as well as several pseudogenes in mouse (Puente et al. 2003; Puente and López-Otín 2004; Pavlopoulou et al. 2010; Koumandou and Scorilas 2013; Lundwall 2013). Moreover, a *KLK1* duplication leading to *KLK2* seems to have occurred early in the mammal tree (Boreoeutheria), however, this appears to have been deleted or silenced into a pseudogene in many species. Finally, in the primate lineage, in a common ancestor of Old Word Monkeys (Catarrhini), a recent duplication of *KLK2* generated *KLK3* (Elliott et al. 2006; Pavlopoulou et al. 2010) (Figure 3).

2.3. Biological importance in human health and disease

Kallikreins are often co-expressed in a wide variety of tissues (Figure 4) and found in many biological fluids. Accordingly, KLKs have been implicated in a broad range of physiological functions and proteolytic cascades, including semen liquefaction, skin desquamation, tooth enamel formation, neural plasticity and regulation of blood pressure (Klokk et al. 2006; Shaw and Diamandis 2007; Lundwall and Brattsand 2008; Lawrence et al. 2010; Prassas et al. 2015). Moreover, the activity of KLKs is generally regulated by fine-tuned processes and spatial-temporal modifications of KLK activity have been associated with pathological conditions such as psoriasis, atopic dermatitis, hypertension, diabetes, neurodegenerative disorders (Alzheimer's and Parkinson's disease) and several types of cancer (Bhoola et al. 1992; Jaffa et al. 1992; Ogawa et al. 2000; Shimizu-Okabe et al. 2001; Sharma 2003; Diamandis et al. 2004; Kontos and Scorilas 2012; Fischer and Meyer-Hoffert 2013; Fuhrman-Luck et al. 2014; Prassas et al. 2015). In this scope, KLK3, also known as the prostate-specific antigen (PSA), is by far the most studied member of the KLK family, given its relevance as a tumor marker for prostate cancer. However, recent studies have shown that other KLKs have been proposed as potential biomarkers mainly because of their deregulation in different types of cancer (Clements et al. 2004; Kontos and Scorilas 2012; Fuhrman-Luck et al. 2014).

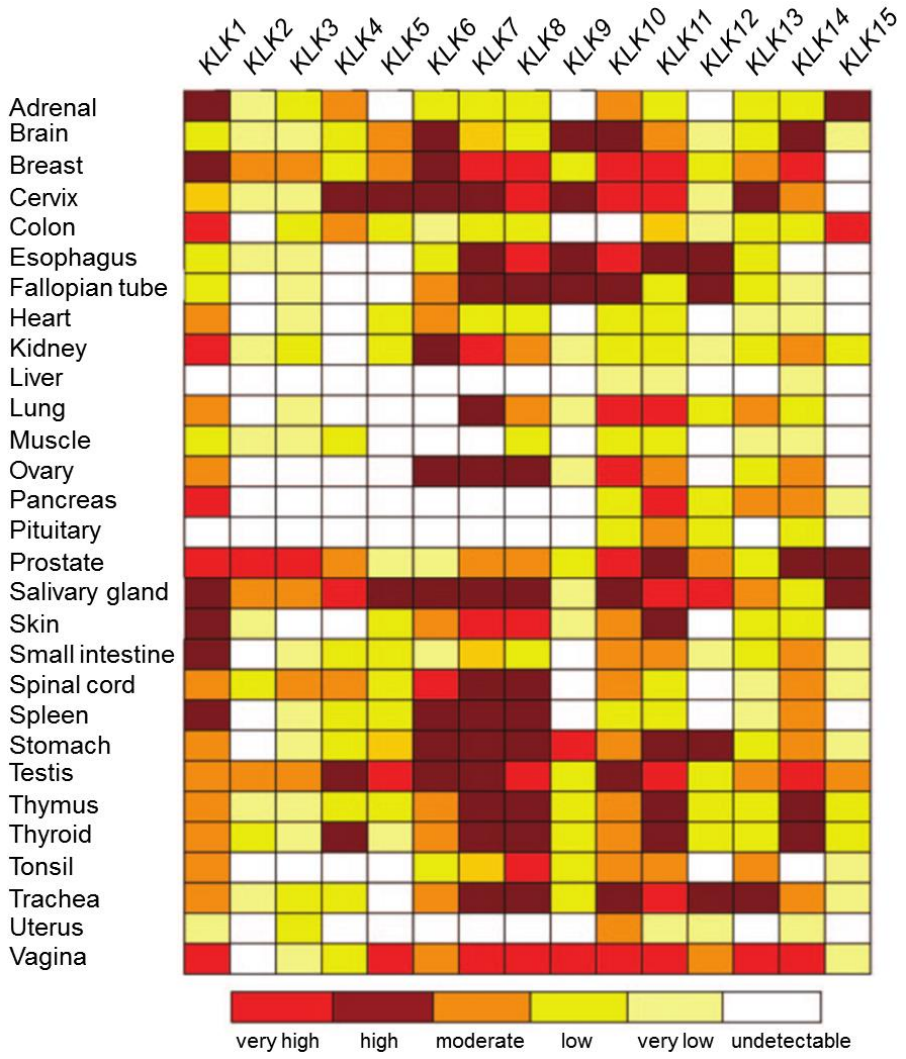


Figure 4 –*KLKs* expression patterns in adult tissues. mRNA concentration for each *KLK* (as indicated in top row) and tissue. The color code at the bottom shows the levels of expression (from Shaw and Diamandis 2007).

2.3.1. Functions in reproductive biology

Semen, also known as seminal fluid, is a complex organic medium produced by male reproductive organs and sex accessory glands. It is composed by a wide range of substances that are essential for proper spermatozoa function and fundamental for semen coagulation and liquefaction processes (Jequier 2000). The epididymal fluid that contains the spermatozoa represents only a small fraction of the total ejaculated volume (<5%), whereas the seminal vesicles secretions, including major structural proteins, account for the largest semen fraction (approximately 65%). The remaining portion of semen volume comes mostly from prostate secretions, in a fluid enriched in proteases and Zn^{2+}

(approximately 30%) (Lundwall and Brattsand 2008). At the moment of ejaculation, in normal physiological conditions, prostate and seminal vesicle secretions are mixed with the epididymal fluid to form a coagulum in the vaginal cavity (Michael et al. 2006; Malm et al. 2007). This gelatinous mass results from the cross-linking of semenogelin 1 (SEMG1), semenogelin 2 (SEMG2) and fibronectin (FN), which are the predominant structural proteins of the seminal plasma, and has the role of entrapping and protecting the spermatozoa. Later, the liquefaction of the coagulum (5-20 minutes after ejaculation) allows a progressive release of the motile spermatozoa together with small peptides that protect sperm cells with their antibacterial, antiviral and antifungal properties (Lilja and Laurell 1985; Lilja et al. 1989; Malm et al. 1996; Peter et al. 1998; de Lamirande et al. 2001; Michael et al. 2006; Edstrom et al. 2008; Lundwall and Brattsand 2008; Zhao et al. 2008).

The liquefaction process (Figure 5), involving a stepwise cleavage of SEMG1, SEMG2 and FN, is essentially driven by KLK3 and KLK2 (Lilja 1985; Lilja et al. 1987; Deperthes et al. 1996). In the prostate, the high concentration of Zn^{2+} regulates the KLK activity with a reversible and allosteric inhibition (Jonsson et al. 2005). Upon ejaculation, the Zn^{2+} is redistributed through the SEMGs, which consequently activates KLK3 and KLK2 allowing the liquefaction of the semen coagulum. However, as SEMGs are hydrolyzed, Zn^{2+} is gradually released and the KLKs activity is downregulated again. This mechanism of negative feedback, controlled by the availability of Zn^{2+}, is fundamental to prevent excessive proteolysis which may damage spermatozoa integrity (Robert and Gagnon 1999; Emami and Diamandis 2007). At this stage, a number of endogenous inhibitors and regulatory feedback loops tightly control the cleaving process. For instance, protein C inhibitor (*SERPINA5*) and eppin (*SPINLW1*) are two serine protease inhibitors that are known to form complexes with SEMGs, preventing a premature proteolytic cleavage of the semen coagulum by KLK3 (Suzuki et al. 2007; Wang et al. 2007; McCrudden et al. 2008).

Importantly, recent studies are proposing that other members of the KLK family beside KLK3 might also play a role in the cascade of semen liquefaction. For example, KLK2, KLK4, KLK5, KLK14 and KLK15 were shown to activate pro-KLK3 *in vitro* (Deperthes et al. 1996; Takayama et al. 2001a; Takayama et al. 2001b; Michael et al. 2006; Emami and Diamandis 2008) and KLK5 and KLK14 were also reported to cleave SEMGs and FN *in vitro* (Michael et al. 2005; Michael et al. 2006; Emami et al. 2008).

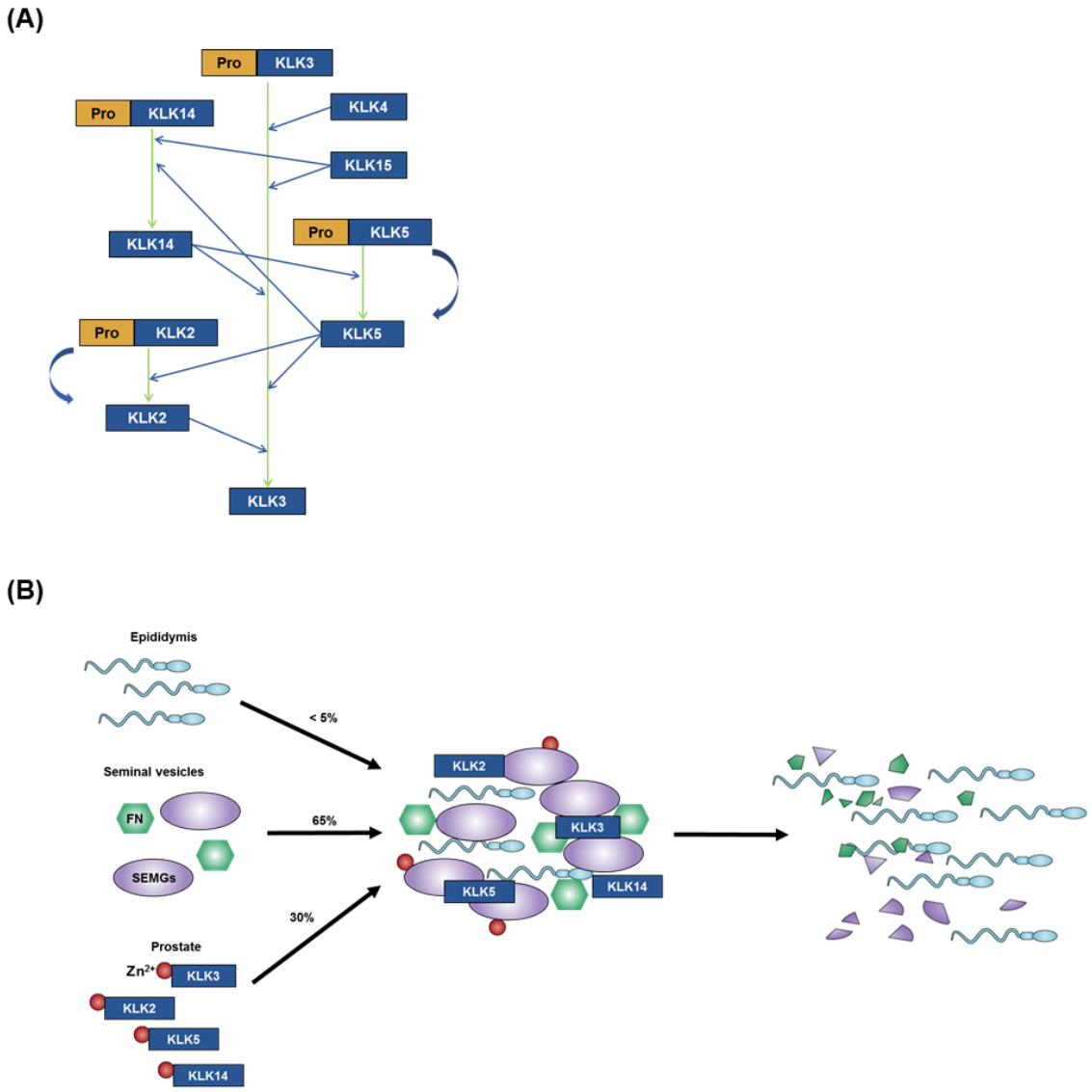


Figure 5 - Schematic representation of the semen liquefaction proteolytic cascade. (A) In normal physiologic conditions, KLKs are activated in the prostate through a zymogen activation cascade. KLK activation by other KLK is represented by straight arrows and auto-activation ability is illustrated by curved arrows. The pro-peptide is represented by the yellow rectangle. **(B)** Upon ejaculation, the sperm-rich epididymal fluid is mixed with prostatic fluids (including KLKs) along with secretions of the seminal vesicles (including SEMG1, SEMG2 and FN), forming the semen coagulum. SEMGs chelate Zn^{2+} ions, which leads to KLK reactivation and subsequent proteolysis of the SEMGs and FN, resulting in seminal coagulum liquefaction (adapted from Michael et al. 2006 and Prassas et al. 2015).

Despite the recent progress in the understanding of KLKs physiological functions in semen liquefaction, their pathological relevance in male infertility remains largely unknown. Currently, it is estimated that in Western countries approximately 15-20% of couples within reproductive age experience difficulties in achieving pregnancy after one year of regular sexual intercourse. Furthermore, among these couples the male factor

contributes with approximately 50% and in about 20% of cases it may inclusively be the single cause of infertility (Organization 1992; Cedenho 2007; Practice Committee of American Society for Reproductive 2012). The male infertility factor is mainly assessed by a semen analysis (spermiogram), where both macro- and microscopic parameters are examined, such as coagulation and liquefaction, viscosity, pH, sperm concentration, morphology and motility (Table 2). Given the importance of KLKs in semen liquefaction and the previous findings of an association between abnormal semen parameters and prostate dysfunction, a possible role of KLKs in infertility was advanced (Andrade-Rocha 2003; Emami et al. 2009; Du Plessis et al. 2013). Supporting this hypothesis is the evidence of aberrant KLK expression found in individuals with abnormal semen liquefaction/viscosity. While a delayed liquefaction was associated with a lower expression of KLK2, KLK3, KLK13 and KLK14, a hyperviscosity phenotype was correlated not only to the down-regulation of KLK2, KLK13 and KLK14 but also KLK1, KLK5-KLK8 and KLK10 (Emami et al. 2009). Moreover, KLK14 expression was found to be significantly reduced in asthenozoospermic patients and, in a recent proteomics study of seminal plasma, KLK3 was found to be upregulated in patients with oligoteratozoospermia (Emami et al. 2009; Sharma et al. 2013).

Table 2 – Semen quality nomenclature according to WHO 1999.

Phenotype	Description
Normozoospermia	Normal ejaculate as defined by reference values
Oligozoospermia	Sperm concentration < 20 x 10^6 spermatozoa/mL
Asthenozoospermia	Percentage of spermatozoa with rapid progressive motility < 25%
Teratozoospermia*	Percentage of morphologically normal spermatozoa < 4%
Oligoasthenozoospermia	Sperm concentration < 20 x 10^6 spermatozoa/mL and percentage of spermatozoa with rapid progressive motility < 25%
Oligoteratozoospermia	Sperm concentration < 20 x 10^6 spermatozoa/mL and percentage of morphologically normal spermatozoa < 4%
Asthenoteratozoospermia	Percentage of spermatozoa with rapid progressive motility < 25% and percentage of morphologically normal spermatozoa < 4%
Oligoasthenoteratozoospermia	Sperm concentration < 20 x 10^6 spermatozoa/mL, percentage of spermatozoa with rapid progressive motility < 25% and percentage of morphologically normal spermatozoa < 4%
Delayed liquefaction	Complete liquefaction does not occur in 60 minutes at 37ºC
Hiperviscosity	Semen forms a thread larger than 2 cm long

* The morphology parameter is defined according to WHO 2010.

2.3.2. Functions in skin physiology

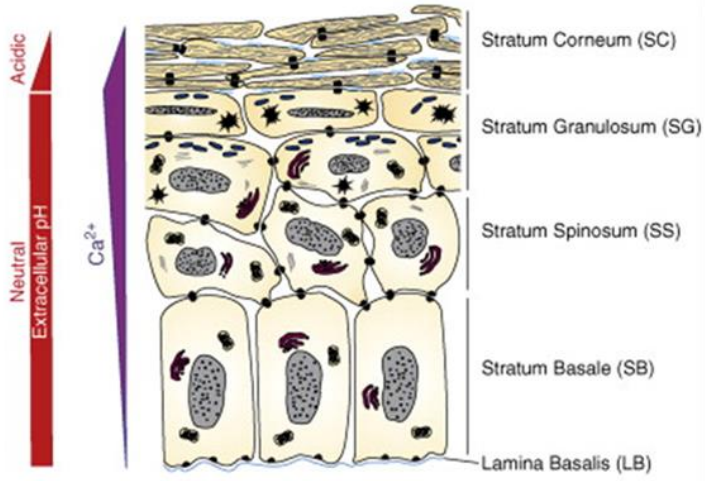
The skin is a multifunctional organ that maintains the body temperature, protects from water loss, mechanical stress, UV-light and also works as a first barrier against infectious agents (bacteria and virus). All these functions of the skin are essentially attributed to the epidermis, the outermost layer which histologically is mainly composed by keratinocytes. These cells are formed in a basal layer (*stratum basale,* SB) and undergo differentiation throughout their migration towards the skin surface, the *stratum corneum* (SC). By the time keratinocytes reach the SC, they have already completed their differentiation into corneocytes and are filled with keratin and are metabolically dead (Figure 6A).

A fine balance between cell proliferation and shedding is crucial to maintain the skin barrier integrity, which is mainly achieved by an active renewal process called skin desquamation (Lundwall and Brattsand 2008; Ishida-Yamamoto et al. 2011; Nishifuji and Yoon 2013; Ishida-Yamamoto and Igawa 2014). In this process, KLK5 and KLK7 play an important role through the proteolytic cleavage of corneodesmosomes, the cell-cell junction structures that mediate corneocyte cohesion. Whereas KLK5 cleaves corneodesmosin (CDSN), desmocollin 1 (DSC1) and desmoglein 1 (DSG1), KLK7 is only able to digest the first two molecules (Caubet et al. 2004).

The majority of the studies regarding KLKs and skin physiology have focused on the importance of KLK5 and KLK7 in skin desquamation but recent findings suggest that other KLKs might also be involved in epidermal pathways. For instance, different levels of KLK1, KLK6, KLK8 and KLK14 were reported in the SC and, except for KLK8, all of them were shown to degrade DSG1 *in vitro* (Borgono et al. 2007; Kishibe 2014; Prassas et al. 2015). Furthermore, KLK5, KLK6, KLK7 and KLK14 were all described as regulated in the skin by the serine protease inhibitor Kazal-type 5 (SPINK5*,* also known as lymphoepithelial kazal type inhibitor or LEKTI) and by alterations in the physiological pH (Caubet et al. 2004; Brattsand et al. 2005; Egelrud et al. 2005; Deraison et al. 2007; Fortugno et al. 2011). In this regard, a complex proteolytic cascade has been proposed to regulate the process of skin desquamation (Figure 6B).

Apart from the functions reported above, KLKs might also be involved in epidermal inflammation, modulation of the lipid-permeability barrier and melanosome transfer, all mediated through the activation of the protease-activated receptor-2 (PAR-2), a transmebrane G-protein-coupled receptor present in keratinocytes. Indeed, PAR-2 receptors were found to be overexpressed in different skin disorders characterized by epidermal barrier defects and inflammation, such as atopic dermatitis (AD), psoriasis and

A)



B)

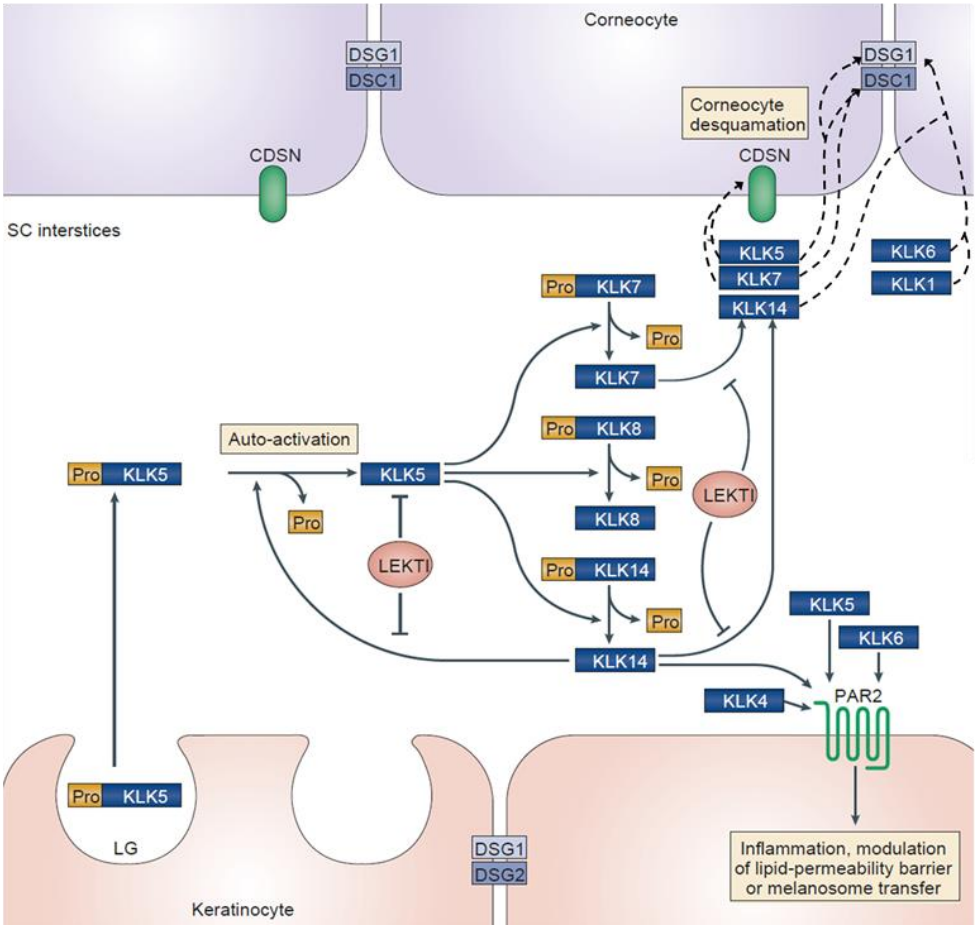


Figure 6 – Schematic representation of epidermis architecture and KLK proteolytic cascade in the skin. (A) The epidermis is organized in different layers mainly arranged by keratinocytes in different stages of differentiation. Keratinocytes are formed in the basal layer (*stratum basale*, SB) and begin to differentiate in the *stratum spinosum* (SS). This differentiation process occurs as keratinocytes migrate towards the skin surface. By the time these cells reach the *stratum corneum* (SC) they have already differentiated into corneocytes, cells filled with keratin and metabolically dead. **(B)** Pro-KLKs are secreted at the SG by lamellar granules (LG) of keratinocytes into SC interstices, where activation occurs by removal of the pro-peptide

Figure 6 (cont.) - (yellow rectangle). Once active, KLKs cleave the corneodesmosome proteins, desmoglein 1 (DSG1), desmocollin 1 (DSC1) and corneodesmosin (CDSN), resulting in corneocyte shedding (skin desquamation). Several KLKs (KLK4, KLK5, KLK6 and KLK14) may also activate the protease-activated receptor-2 (PAR-2), leading to inflammation, modulation of lipid-permeability barrier or melanosome transfer. The KLK activity in the skin is regulated by protease inhibitors, such as serine protease inhibitor Kazal-type 5 (SPINK5 or LEKTI), and by the epidermal pH gradient (adapted from Ovaere et al. 2009 and Prassas et al. 2015).

Netherton syndrome (NS) (Buddenkotte et al. 2005; Descargues et al. 2006; Eissa and Diamandis 2008; Lee et al. 2010). Furthermore, the similar localization pattern of *PAR-2* and several *KLKs* (*KLK1*, *KLK4-KLK7*, *KLK9-KLK11*, *KLK13-KLK14*) in skin lesions of AD and NS patients, together with the *in vitro* activation of PAR-2 by KLK4, KLK5, KLK6 and KLK14, seems to support the existence of a KLK-PAR-2 activation mechanism (Steinhoff et al. 2003; Komatsu et al. 2005; Descargues et al. 2006; Komatsu et al. 2007a). In addition, the deregulation of KLK activity has been thought to be one of the triggering causes of AD, NS and psoriasis, moreover, several clinical studies have already demonstrated the upregulation of multiple KLKs in these pathologies (Komatsu et al. 2002; Komatsu et al. 2005; Komatsu et al. 2007a; Komatsu et al. 2007b; Komatsu et al. 2008). In the NS, *SPINK5* disruptive mutations have been reported, so far, as the single genetic cause underlying abnormal KLK activity (Chavanas et al. 2000; Walley et al. 2001; Bitoun et al. 2002; Komatsu et al. 2002; Komatsu et al. 2008). Consistently, *Spink5* knockout mouse model was found to display an uncontrolled Klk activity, as observed in patients with NS (Descargues et al. 2005).

2.3.3. Functions in tooth enamel formation

Dental enamel is the most highly mineralized tissue in the human body and its developmental stage (amelogenesis) is tightly controlled by two major proteases, MMP20 and KLK4 (Bartlett and Simmer 1999; Simmer and Hu 2002; Lu et al. 2008). These two proteases are responsible for the hydrolysis of the dental extracellular matrix proteins: amelogenin, ameloblastin and enamelin, which account for about 90%, 5% and 2% of the total of the enamel protein, respectively (do Espirito Santo and Line 2005). Briefly, amelogenesis can be divided into three major stages: the secretory, the transitional and the maturation stage. In the first two stages, MMP20 is the predominantly expressed protease and it is responsible for the early processing of amelogenin and ameloblastin, allowing enamel crystallites to grow in length (Ryu et al. 1999; Iwata et al. 2007; Lu et al.

2008; Nagano et al. 2009). On the other hand, KLK4 expression is only initiated in the transitional stage and continues throughout the maturation phase, further contributing to the degradation of matrix proteins. Specifically, KLK4 favors crystal growth in width and thickness, thus promoting enamel hardening (Figure 7) (Hu et al. 2002; Ryu et al. 2002; Simmer and Hu 2002; Yamakoshi et al. 2006; Lu et al. 2008; Zhu et al. 2014).

The characterization of MMP20 and KLK4 activities in the different stages of amelogenesis was only possible through the use of knockout mouse models lacking either *Klk4* or *Mmp20* (Yamakoshi et al. 2011). While non-cleaved amelogenin and ameloblastin were only observed in the *Mmp20* null mice in the secretory stage, in the *Klk4* null mice only enamel proteins were accumulated in late maturation phase (Yamakoshi et al. 2011). Furthermore, whereas mice lacking *Klk4* produced enamel of normal thickness but with softer inner regions, the *Mmp20* null mice presented thin and structurally abnormal enamel (Simmer et al. 2009; Smith et al. 2011). Importantly, KLK4 was described as a pro-enzyme activated by MMP20 and DPPI based on *in vitro* experiments (Ryu et al. 2002; Tye et al. 2009; Yamakoshi et al. 2013). However, *KLK4* and *MMP20* expression overlap only during a short time period in the transitional stage, on the contrary, *DPP1* expression is maintained throughout enamel formation, suggesting this as the predominant activator of KLK4 (Hu et al. 2002; Tye et al. 2009).

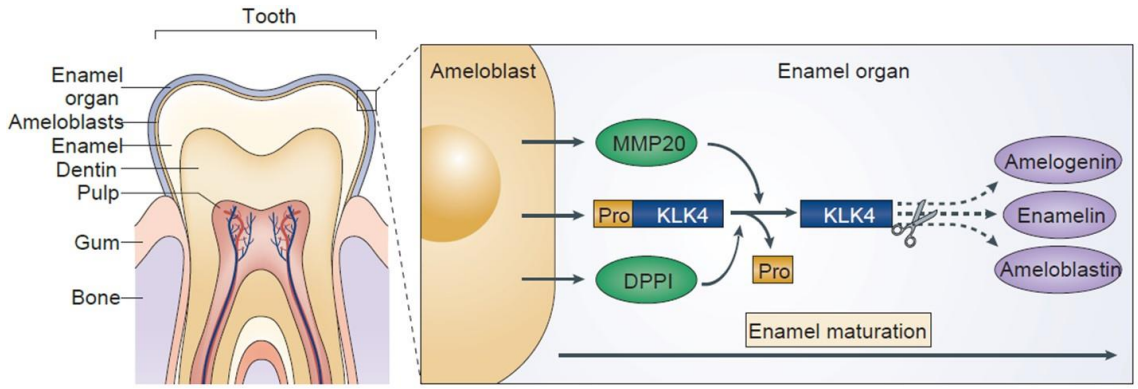


Figure 7 – KLK4 in tooth enamel formation. The ameloblasts secrete a protein-rich matrix composed by amelogenin, enamelin and ameloblastin, as well as KLK4, MMP20 and DPP1 proteases. During the transitional and maturation stages, pro-KLK4 is secreted and activated by MMP20 and DPP1. Upon activation, KLK4 degrades the dental extracellular matrix proteins, allowing crystal growth in width and thickness, thus promoting enamel hardening (adapted from Prassas et al. 2015).

In the amelogenesis framework, the importance of KLK4 is further highlighted by disruptive mutations in the human *KLK4* gene (W153X and G82Afs*87), which cause specific tooth malformations, such as enamel hypomaturation, tooth hyper-pigmentation and hyper-sensitivity to hot and cold stimuli (Hart et al. 2004; Wright et al. 2006; Wright et

al. 2011; Wang et al. 2013). This genetic condition, known as hypomaturation-type *amelogenesis imperfecta* (AI2A1), has been reported in three unrelated families, two carrying the W153X mutation and one carrying the G82Afs*87. In all cases, the affected patients were homozygous for the loss-of-function mutations (Hart et al. 2004; Wright et al. 2011; Wang et al. 2013).

2.4. Primate adaptive evolution

Primate mating behavior was found to drive the intensity of sperm competition and, concomitantly, the evolution of genes involved in reproduction. Specifically, promiscuous species in which females are known to mate with more than one male per periovular period (polyandrous or multimale/multifemale) exhibit physiological traits better adapted for fertilization, like larger testis and larger seminal vesicles relative to body size (Harcourt et al. 1981; Dixson and Anderson 2002). Another feature that was shown to correlate to mating system and promiscuity was the thickness of the seminal coagulum or the presence of a copulatory plug (Dixson and Anderson 2002). Broadly speaking, a thicker coagulum is more efficient in preventing the fertilization, by rival males, of a recently inseminated female in succeeding copulations than a gelatinous one (Jensen-Seaman and Li 2003). In this context, SEMGs represent one of the best known examples of evolution driven by post-copulatory selection, in which higher d_N/d_S ratio values observed in chimpanzees (*Pan troglodytes*) were correlated to larger numbers of male partners per periovular period and increased thickness of the semen coagulum. In contrast, in monoandrous taxa, the semen coagulum is fluid and SEMGs seem to have undergone a more relaxed evolution, accumulating of several inactivating mutations (Jensen-Seaman and Li 2003; Dorus et al. 2004; Hurle et al. 2007).

Considering the role of KLKs in the hydrolysis of SEMGs, it is likely that KLKs evolution might also correlate to sperm competition. In fact, loss of KLK2, either by gene deletion or inactivating mutations, has already been reported in several species with different outcomes in reproductive biology (Clark and Swanson 2005). On one hand, the absence of *KLK2* was described in gorillas (*Gorilla gorilla*) and gibbons (Hylobates sp.), two examples of monoandrous species, in which *SEMG1* and *SEMG2* have been inactivated due to the accumulation of premature stop codons. In these species, the loss of SEMGs are thought to diminish the semen viscosity, impairing the formation of the semen coagulum and possibly rendering *KLK2* activity redundant (Clark and Swanson 2005). On the other hand, in the rhesus monkey (*Macaca mulatta*), the presence of a

KLK2 mutation affecting the catalytic triad (D102A) and, consequently, enzyme activity was proposed to contribute to the different semen physiology of this species, in which the semen does not liquefy but forms a copulatory plug instead (Clark and Swanson 2005). Given that the rhesus monkey is a polygamous species in nature, the presence of a copulatory plug is important for sperm competition and mate guarding. However, taking into account that *SEMG1* has been inactivated by a frameshift mutation in this species too, *KLK2* loss-of-function could result in a reduced ability to dissolve the semen coagulum, thus allowing the formation of a copulatory plug. Still, the role of *KLKs* in the reproductive system and the dynamical implications of mating behavior remain poorly characterized.

Chapter 2

Aims

The importance of *KLK*s in the cascade of semen liquefaction together with the evidence that proteolytic and reproductive genes might have been targeted by natural selection, at inter- and intraspecific level, has motivated the characterization of *KLK* sequence variation in different primate species, and in healthy and diseased human populations.

Specific aims of this work:

1. Unravel the evolutionary history of the most recent *KLK* duplicates, *KLK2* and *KLK3*, and address a possible correlation to primate mating systems and sperm competition. In a systematic analysis of *KLK2* and *KLK3,* using comparative and phylogenetic procedures, a total of 22 primate species with diverse mating systems and different patterns of semen coagulation were investigated.

2. Characterize a signature of natural selection shaping *KLK* cluster diversity in Asian populations. The hypothesis of a potential signature of natural selection among *KLK* genes in Asians, as previously highlighted by independent genome-wide scans of positive selection, was accomplished by undertaking a comprehensive survey of 1000 Genomes (phase I) data and combining it with *in vitro* functional assays for the most likely candidate variants. Furthermore, a possible ascertainment bias of the 1000 Genomes data was evaluated by Sanger sequencing of several genomic segments across the *KLK* cluster in a subsample of individuals screened by the 1000 Genomes project.

3. Assess the impact of *KLK* sequence variation in different infertility phenotypes. An association study of male infertility centered in the genetic screening of the *KLK* cluster, along with their targets, *SEMG1* and *SEMG2*, and their potential inhibitors of the whey acidic protein four-disulfide core domain (*WFDC*) locus was performed in a cohort of Portuguese infertility cases and controls. The survey for potential candidate variants for male infertility was achieved through a combination of pooled sample high-throughput sequencing, Sanger sequencing and other genotype screening methods.

Chapter 3

Papers

Paper I - Birth-and-Death of *KLK3* and *KLK2* in Primates: Evolution Driven by Reproductive Biology

Genome Biol. Evol. 2012 4(12):1331-1338

GBE

Birth-and-Death of *KLK3* and *KLK2* in Primates: Evolution Driven by Reproductive Biology

Patrícia Isabel Marques[1,2,3], Rui Bernardino[4], Teresa Fernandes[4], NISC Comparative Sequencing Program[5,6], Eric D. Green[5], Belen Hurle[5], Victor Quesada[2,*], and Susana Seixas[1,*]

[1]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal

[2]Department of Biochemistry and Molecular Biology-IUOPA, University of Oviedo, Oviedo, Spain

[3]Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal

[4]Lisbon Zoo Veterinary Hospital, Lisbon, Portugal

[5]National Human Genome Research Institute, National Institutes of Health (NIH), Bethesda, Maryland

[6]NIH Intramural Sequencing Center (NISC), National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland

*Corresponding author: E-mail: sseixas@ipatimup.pt; quesadavictor@uniovi.es.

Accepted: November 25, 2012

Data deposition: GenBank accession numbers for all the BAC genomic entries and the assembly coordinates of the genomic segments from reference sequences (UCSC Genome Browser) used in gene annotation are provided in supplementary table S1, Supplementary Material online.

Abstract

The *kallikrein* (*KLK*) gene family comprises the largest uninterrupted locus of serine proteases in the human genome and represents a notable case for studying the evolutionary fate of duplicated genes. In primates, a recent duplication event gave rise to *KLK2* and *KLK3*, both encoding essential proteins for the cascade of seminal plasma liquefaction. We reconstructed the evolutionary history of *KLK2* and *KLK3* by comparative analysis of the orthologous sequences from 22 primate species, calculated d_N/d_S ratios, and addressed the hypothesis of coevolution with their substrates, the semenogelins (SEMG1 and SEMG2). Our findings support the placement of the *KLK2–KLK3* duplication in the Catarrhini ancestor and unveil the frequent loss of *KLK2* throughout primate evolution by different genomic mechanisms, including unequal crossing-over, deletions, and pseudogenization. We provide evidences for an adaptive evolution of KLK3 toward an expanded enzymatic spectrum, with an effect on the hydrolysis of semen coagulum. Furthermore, we found associations between mating system, the number of SEMG repeat units, and the number of functional *KLK2* and *KLK3*, suggesting complex evolutionary dynamics shaped by reproductive biology.

Key words: serine proteases, adaptive evolution, mating system, semen coagulation, semenogelins.

The birth-and-death of genes has a significant impact in genome evolution, particularly in gene families involved in physiological traits such as sensory systems, immunity, and reproduction (Zhang 2003; Demuth and Hahn 2009). According to this model, new genes are generated by duplication, and although some are maintained in the genome (acquiring novel or altered functions), others are disrupted or become nonfunctional through a variety of deleterious mechanisms (Nei and Rooney 2005; Kaessmann 2010). In this context of gene gain, diversification, and loss, the *kallikrein* (*KLK*) cluster, the largest locus in the human genome of phylogenetically related serine proteases (Yousef and Diamandis 2001), represents a remarkable case for the study of the evolutionary fate of duplicates. In humans, the *KLK* cluster spans over 265 kb on chromosome 19q13.4 and includes 15 genes ranging from 4.4 to 10.5 kb, most of them sharing a common gene structure with five coding exons (Yousef and Diamandis 2001; Lundwall and Brattsand 2008). KLKs act mainly as trypsin or chymotrypsin-like proteases in a number of biological processes such as skin desquamation, semen liquefaction, neuroplasticity, and regulation of blood pressure (Emami and Diamandis 2007). In primates, a recent gene duplication gave rise to two kallikreins, *KLK2* and *KLK3* (encoding prostate-specific antigen), which play a crucial role in the proteolytic cascade of seminal plasma liquefaction (Lundwall and Brattsand 2008). Briefly, upon ejaculation, the epididymal fluid is mixed with prostate and seminal vesicles secretions containing semenogelins

(SEMG1 and SEMG2) to form a coagulum that entraps spermatozoa. Later, these spermatozoa are released with the hydrolysis of SEMGs by KLK3 and KLK2. In addition, KLK2 is also thought to activate KLK3 (Lovgren et al. 1999; Lundwall and Brattsand 2008). Previous findings suggest that primate *KLK2* and *KLK3* (Clark and Swanson 2005), along with *SEMGs* (Jensen-Seaman and Li 2003; Dorus et al. 2004; Hurle et al. 2007), may be targets of natural selection and could provide an important example of birth-and-death evolution. Here, we reconstruct the evolutionary history of *KLK2* and *KLK3* in primates and test the hypothesis of their coevolution with *SEMGs* as a possible example of evolution driven by male reproductive biology.

KLK2 and *KLK3* Gains and Losses

To better understand the evolutionary dynamics of *KLK2* and *KLK3* genes in primates, we sequenced and/or annotated the orthologous genomic segments spanning these genes in a total of 22 primate species (supplementary table S1, Supplementary Material online). We confirmed the presence of *KLK2* and *KLK3* in all Catarrhini, except for *Colobus guereza*, *Gorilla gorilla,* and *Nomascus leucogenys*, and the presence of a single *KLK2* ortholog sequence in Platyrrhini and Strepsirrhini (fig. 1*A*). This result reinforces the hypothesis of a *KLK3* origin by *KLK2* duplication after the Catarrhini split approximately 42 million years ago (Olsson et al. 2004;

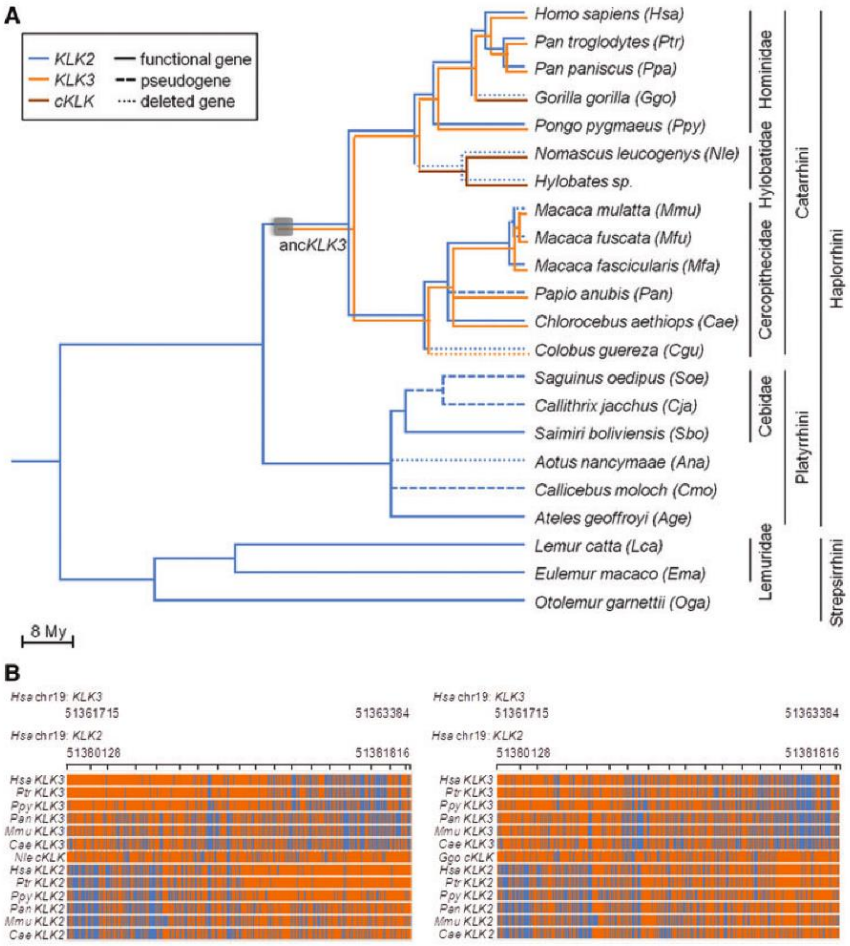


Fig. 1.—Phylogenetic analysis of *KLK2* and *KLK3* in primates. (*A*) Phylogenetic tree showing primate divergence times (Hedges et al. 2006) and functional status of *KLK2* and *KLK3*. The criteria to define a nonfunctional *KLK* gene were the identification of at least one disrupting mutation. Gray square indicates a duplication event. The ancestral *KLK3* branch is indicated (anc*KLK3*). (*B*) Alignment of exons IV–V for *KLK2* and *KLK3* in Catarrhini. The corresponding human genomic positions for these regions are represented at the top. Positions conserved with *Gorilla gorilla* (left panel) or *Nomascus leucogenys* (right panel) are in orange. Nonconserved positions are in blue. Sites conserved in all species were omitted.

Table 1
Identified *KLK2* Deleterious Mutations

Species	Deleterious Mutations[a]
Hsa, Ptr, Ppa, Ppy, Mfa, Cae, Sbo, Age, Lca, Ema, Oga	None
Mmu	D120A[b,c] and R109X[d]
Mfu	G51R,[c,d] L54P,[c,d] and L171fsX181[d]
Pan	V247fsX337
Cja	M1I, W47X, IVS3+1G>A,[e] IVS3-1G>A,[e] and L178_C184delinsVfsX190
Soe	M1I,[f] S213T,[c,f] R250X,[f] and IVS4+1G>A[f]
Cmo	M1L, I25T, C184fsX189, and IVS4-2A>T[e]

[a]Mutations are displayed according to the recommended nomenclature for the description of human sequence variations (den Dunnen and Antonarakis 2000).
[b]Catalytic triad mutation (Clark and Swanson 2005).
[c]Possible damaging as predicted by Polyphen2 (Adzhubei et al. 2010).
[d]Polymorphic site.
[e]Splice site mutation.
[f]Previously identified mutations (Olsson et al. 2004).

Valtonen-Andre et al. 2005; Pavlopoulou et al. 2010). Notably, we identified two *KLK3–KLK2* fusions in *G. gorilla* and *N. leucogenys* yielding single chimeric *KLK* genes (*cKLK*) (supplementary fig. S1*A* and S1*B*, Supplementary Material online). We located the breaking point in both species to a few bases in intron IV, in the vicinity of a LINE2 element common to *KLK2* and *KLK3* sequences (*G. gorilla* IVS4+622_781 and *N. leucogenys* IVS4+787_1026; fig. 1*B*). These genomic rearrangements were confirmed by direct sequencing of three additional *G. gorilla* individuals and five Hylobatidae samples, indicating a likely fixation of *cKLK* in these taxa (supplementary fig. S1*C*, Supplementary Material online). In both cases, the first four exons of *cKLK* are orthologous to *KLK3*, whereas the last exon is more similar to *KLK2* (fig. 1*B*). At the protein level, these genomic rearrangements account only for minor amino acid replacements relative to the expected KLK3 sequence (S231P, R239K, S241A, L242V, and V258A). Because these replacements are not predicted to alter protein structure or function, *cKLK* is likely a functional *KLK3*-like gene. Our findings confirm previous reports, which suggested the partial loss of *KLK2* in *G. gorilla* and *Hylobates* sp. (Clark and Swanson 2005). On the other hand, a detailed analysis of the alignments of *C. guereza* genomic sequences with the *Homo sapiens* reference genome showed the complete loss of *KLK2* and *KLK3* in this species, possibly by two deletion events (supplementary fig. S2, Supplementary Material online). In Cercopithecoidea, we identified several loss-of-function events in *KLK2* through a variety of deleterious mechanisms (table 1). These include a premature stop codon in *Macaca mulatta* (R109X) and a frameshift mutation (L171fsX181) and two nonsynonymous substitutions (G51R and L54P) in *M. fuscata*. The mutation of the KLK2 catalytic triad (D120A) previously described in *M. mulatta* (Clark and Swanson 2005) was not observed, and no evidence for the accumulation of deleterious mutations in *M. fascicularis* was found. In *Papio anubis*, we identified a frameshift mutation leading to a 75-codon longer open reading frame (V247fsX337), which is unlikely to be translated into a KLK2 (supplementary fig. S3, Supplementary Material online). Additional examples of *KLK2* loss were observed in Platyrrhini, either by gene deletion or disruption (fig. 1*A* and table 1). In this taxon, the single example of *KLK2* loss by deletion was found in *Aotus nancymaae*, whereas several deleterious mutations were detected in *Callicebus moloch*, *Callithrix jacchus*, and *Saguinus oedipus*. In *Cal. moloch*, these mutations affect the starting codon (ATG-TTG), alter the activation site (I25T), and produce a premature stop codon (C184fsX189). In *Callithrix jacchus*, we identified a disrupted start codon (ATG-ATA) and a premature stop codon (W47X). In *S. oedipus*, a sister species of *Callithrix jacchus*, we have confirmed that *KLK2* is a pseudogene due to the accumulation of several mutations predicted to impair the translation of a functional serine protease (Olsson et al. 2004). All these species have an alternative starting codon 18 bp upstream of the consensus site; however, this is not expected to lead to an active KLK2 due to the occurrence of additional damaging mutations (supplementary fig. S3, Supplementary Material online, and table 1). In Strepsirrhini, no deleterious mutations were detected, suggesting a functional *KLK2* (fig. 1*A* and supplementary fig. S3, Supplementary Material online).

KLK2 and *KLK3* Phylogenetic Analysis

To address the extent of the selective pressures exerted on *KLK2* and *KLK3*, we calculated d_N/d_S (ω; d_S—synonymous substitution rate and d_N—nonsynonymous substitution rate) ratios under alternative models of gene evolution. To this end, we performed a series of branch models to test whether *KLK2* and *KLK3* experienced different selective pressures during primate evolution. First, we estimated a single ω for the entire phylogeny (one-ratio model), in which we assumed no differentiation in *KLK2* and *KLK3* selective constrains. The observed ω value below 1 ($\omega_{KLK} = 0.54$) pointed out to an overall conservation of *KLK2* and *KLK3* (table 2). Then, to examine whether the two paralogs were subjected to

Table 2
Parameter Estimates and Likelihood Scores under Different Branch Models

Model	Parameters for Branches	Likelihood (*l*)
One ratio	$\omega_{KLK} = 0.54$	−4,390.93
Two ratios	$\omega_{KLK2} = 0.55$	−4,390.90
	$\omega_{KLK3} = 0.53$	
Three ratios	$\omega_{KLK2} = 0.48$	−4,386.40
	$\omega_{pKLK2} = 1.16$	
	$\omega_{KLK3} = 0.53$	
Four ratios	$\omega_{KLK2} = 0.47$	−4,384.35
	$\omega_{pKLK2} = 1.16$	
	$\omega_{KLK3} = 0.43$	
	$\omega_{ancKLK3} = 0.90$	

Models Compared	$-2\Delta l$	*P*
One vs. two ratios	0.06 (df = 1)	0.806
Two vs. three ratios	9.00** (df = 1)	0.003
Three vs. four ratios	4.10* (df = 1)	0.043

NOTE.—ω_{KLK}, ω for all *KLK2* and *KLK3* lineages; ω_{KLK2}, ω for all *KLK2* lineages; ω_{KLK3}, ω for all *KLK3* lineages; ω_{pKLK2}, ω for *KLK2* pseudogene lineages; $\omega_{ancKLK3}$, ω for the ancestral *KLK3* lineage; df - degrees of freedom.

*Significant $P < 0.05$.

**Significant $P < 0.01$.

different selective pressures, we applied a different model (two-ratio model) considering two branches within the phylogeny comprising either the *KLK2* or *KLK3* clades. Both ω values were below 1 ($\omega_{KLK2} = 0.55$; $\omega_{KLK3} = 0.53$) (table 2) and did not differ from the previous reported model ($-2\Delta l = 0.06$; $P > 0.05$). Given the evidences for *KLK2* pseudogenization in several primate species, we anticipated a contrast in the relaxation of selective constraints in pseudogenes and their functional orthologs. In our models, we considered this hypothesis by subdividing *KLK2* clade into functional and pseudogenized (*pKLK2*). This model (three-ratio model) had a significant higher likelihood and an improved fit to *KLK2* and *KLK3* evolution ($-2\Delta l = 9$, $P < 0.01$). Furthermore, ω estimates corroborated the neutral evolution of *KLK2* pseudogenes ($\omega_{pKLK2} = 1.16$), with a possible relaxation of selective constrains after the duplication event ($\omega_{KLK2} = 0.48$; $\omega_{KLK3} = 0.53$; table 2). Therefore, to test whether *KLK3* had been subjected to different selective pressures following the duplication, the ancestral *KLK3* branch (anc*KLK3*) was regarded as an independent clade. The last model (four-ratio model) provided the best fit to the evolutionary history of *KLK2* and *KLK3* ($-2\Delta l = 4.10$, $P < 0.05$). According to the ω values estimated, an episode of reduced selective constrains occurred immediately after the duplication event ($\omega_{ancKLK3} = 0.90$); stronger selective pressures are operating at *KLK3* and *KLK2* ($\omega_{KLK3} = 0.43$; $\omega_{KLK2} = 0.47$), and a complete release of selective constraints is observed among *KLK2* pseudogenes ($\omega_{pKLK2} = 1.16$; table 2). Importantly, if the ancestral *KLK3* experienced brief episodes of adaptive evolution, it is unlikely to produce an ω value greater than 1, because most residues were subjected to strong constrains and only a few were under positive selection. To test the adaptive hypothesis of the ancestral *KLK3*, we performed a branch-site model, in which branches on the phylogeny are divided a priori into foreground (ancestral *KLK3* branch) and background and selective pressures are allowed to vary over sites and branches. Even though the majority of sites are constrained or neutrally evolving, four codon positions (13, 41, 72, and 207) show a footprint of positive selection with posterior probability higher than 85% (table 3). A similar approach was applied to the functional *KLK2* and *KLK3* data sets using the site models test. In these cases, variable ω ratios among sites were calculated for each gene and neutral and selection models compared (M1 vs. M2 and M7 vs. M8). In both cases, selection models fit significantly better the *KLK2* and *KLK3* data than neutral models (table 3), and eight (18, 67, 69, 109, 177, 205, 210, and 250) and five (45, 189, 203, 238, and 248) codon positions were identified as being positively selected in *KLK2* and *KLK3*, respectively (table 3).

To uncover the adaptive impact of the amino acids replacements targeted by positive selection, we mapped the corresponding residues onto three-dimensional models of KLK2 and KLK3. From the eight sites identified for KLK2, amino acids 177 and 210 are located in the catalytic pocket and 109 in the kallikrein loop (fig. 2*A*). Among KLK3 selected sites, the D207S replacement that occurred shortly after the duplication is located at the base of the substrate-biding pocket (fig. 2*B*). Noteworthy, D207S altered enzyme activity to a chymotrypsin-like specificity and modified substrate affinity to medium size hydrophobic (tyrosine, leucine, valine, and phenylalanine) or basic residues (arginine, lysine, and histidine) (Debela et al. 2006). On the other hand, KLK2 conserved the aspartate residue at position 207, which is known to confer trypsin-like specificity to kallikrein-related peptidases and to display a strong preference for arginine in substrates (Janssen et al. 2004; Debela et al. 2006; Emami and Diamandis 2007). The findings of KLK3 adaptive evolution and of its expanded enzymatic spectrum on top of the restricted KLK2 spectrum provide strong arguments for a significant impact of *KLK3* emergence in the hydrolysis of semen coagulum and in the extensive SEMGs fragmentation as currently seen in humans.

Implications of *KLK2* and *KLK3* Evolution in Primate Reproductive Biology

Primate mating behavior drives the intensity of sperm competition and, with that, the evolution of genes involved in reproduction. Specifically, polyandrous species exhibit physiological traits better adapted for fertilization, like larger testis relative to body size (Harcourt et al. 1981; Clark and Swanson 2005). For *SEMGs*, a relationship between molecular evolution rates and female promiscuity has been already shown. The SEMGs are

Table 3
Model Comparisons of Variable ω Ratios among Sites

Models Compared	$-2\Delta l$	Parameter Estimates under Selection	Positively Selected Sites[a]
KLK2			
M1 vs. M2	6.56* (df = 2)	$p_1 = 0.69$, $\omega = 0.16$ $p_2 = 0$, $\omega = 1.00$ $p_3 = 0.31$, $\omega = 1.65$	109
M7 vs. M8	9.78** (df = 2)	$p_0 = 0.69$, $p = 19.76$ $q = 99.00$ ($p^1 = 0.31$), $\omega = 1.66$	18, 67, 69, 109, 177, 205, **210**, 250
KLK3			
M1 vs. M2	5.59 (df = 2)	$p_1 = 0.73$, $\omega = 0$ $p_2 = 0$, $\omega = 1.00$ $p_3 = 0.27$, $\omega = 1.83$	45, **189**
M7 vs. M8	6.37* (df = 2)	$p_0 = 0.73$, $p = 0.01$ $q = 2.82$, ($p^1 = 0.27$), $\omega = 1.83$	**45**, **189**, 203, 238, 248
Branch-site (MA)	13.06**	$p_0 = 0.51$, $\omega_{bg} = 0.13$, $\omega_{fg} = 0.13$ $p_1 = 0.41$, $\omega_{bg} = 1.00$, $\omega_{fg} = 1.00$ $p_{2a} = 0.04$, $\omega_{bg} = 0.13$, $\omega_{fg} = 17.06$ $p_{2b} = 0.03$, $\omega_{bg} = 1.00$, $\omega_{fg} = 17.06$	13, 41, 72, **207**

NOTE.—ω_{bg}, ω for background branches; ω_{fg}, ω for foreground branch (ancestral *KLK3* branch).
[a]Sites with posterior probabilities > 0.85 are indicated in regular type; *P* values > 0.90 are underlined and *P* values > 0.95 are in bold.
*Significant $P < 0.05$.
**Significant $P < 0.01$.

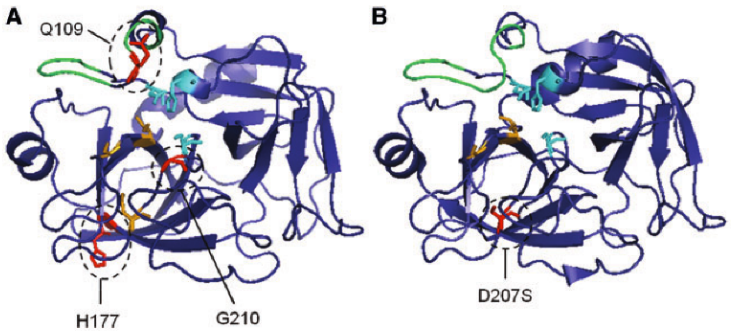


FIG. 2.—Positive selected sites in biologically relevant regions. (*A*) Human KLK2 three-dimensional model showing amino acid replacements predicted to be under positive selection (Q109, H177, and G210). (*B*) Human KLK3 three-dimensional model showing D207S substitution predicted to be under positive selection in the ancestral branch. The catalytic triad is represented in light blue (H65, D120, and S213) and the binding sites in orange (S228, G230, and D207 in KLK2 or S207 in KLK3).

highly polymorphic modular proteins, with a number of repeat units varying within and between species. This in turn dictates the degree of crosslinking between SEMGs, which influences the semen coagulum thickness. The correlation is such that the higher the promiscuity of a given species, the higher the likelihood of longer SEMGs, more crosslinking events, and a rigid copulatory plug, possibly influencing the fertilization of a recently inseminated female by rival males (Jensen-Seaman and Li 2003; Dorus et al. 2004; Hurle et al. 2007). Indeed, the number of SEMG repeats shows a significant rank correlation with residual testis size, which is a good proxy for primate mating system (Anderson et al. 2004; Dixson and Anderson 2004; Wlasiuk and Nachman 2010) (fig. 3*A* and supplementary table S2, Supplementary Material online). Considering the role of KLK2 and KLK3 in the hydrolysis of SEMGs (Rawlings et al. 2012), we tested whether the presence or absence of functional genes correlates with the number of SEMG1 and SEMG2 repeat units. In most cases, active KLK2 and KLK3 are associated with higher repeat numbers and polyandry, whereas the lack of one or both of them is linked to lower repeat numbers and monoandry (fig. 3*B* and *C* and supplementary table S2, Supplementary Material online).

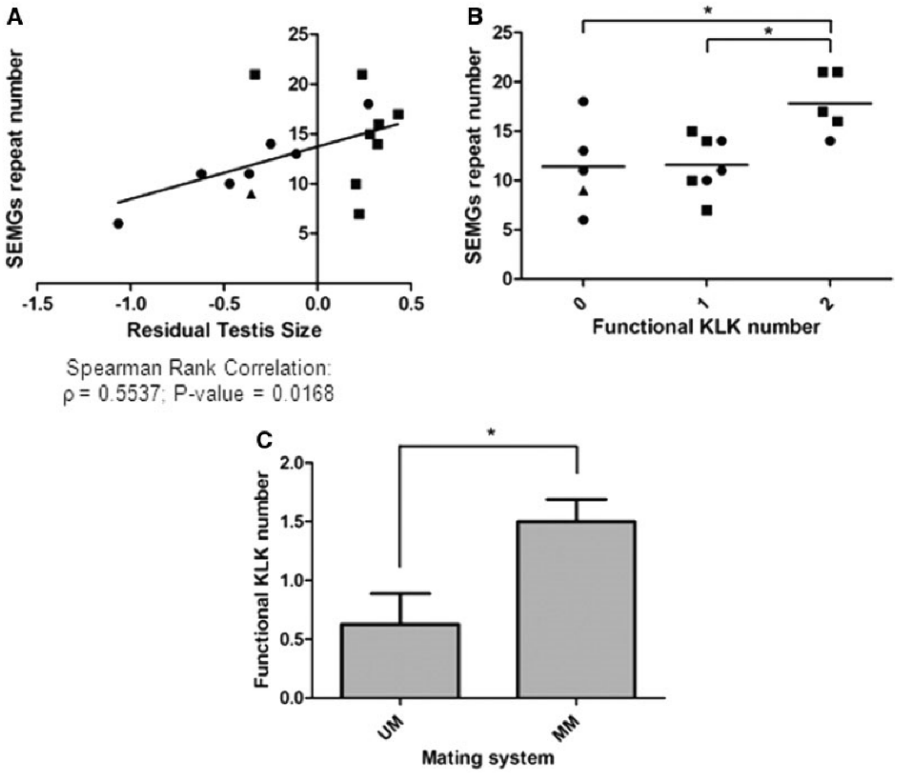


FIG. 3.—Evolution of primate KLK2 and KLK3 related to mating factors. (*A*) Correlation of residual testis size (Anderson et al. 2004; Dixson and Anderson 2004; Wlasiuk and Nachman 2010) with the combined SEMG repeat units (Jensen-Seaman and Li 2003; Hurle et al. 2007). (*B*) Correlation between the number of SEMG1 and SEMG2 repeat units (Jensen-Seaman and Li 2003; Hurle et al. 2007) and the presence of functional KLK2 and KLK3. $*P < 0.05$. (*C*) Correlation between the mating system (Wlasiuk and Nachman 2010) and the presence of functional KLK2 and KLK3. UM, unimale; MM, multimale. $*P < 0.05$. (●), monoandrous; (■), polyandrous; and (▲), ambiguous.

Interestingly, we also observed a trend for higher KLK numbers with more prominent semen coagulation and increase residual testis size (supplementary fig. S4, Supplementary Material online).

We propose a model in which *KLK2* and *KLK3* coevolved with *SEMGs* in a sperm competition-driven process. In a polyandrous species with many SEMG repeats and prominent semen coagulation, the robust and orchestrated activity of KLK2 and KLK3 may be important for sperm release. Conversely, in a monoandrous species with few SEMG repeats, the loss of *KLK2* might represent a biological response to maintain a gelatinous coagulum. Here, the loss of *KLK2* may not be arbitrary because KLK3 has a larger spectrum of cleavage sites in SEMGs than KLK2 (Rawlings et al. 2012), therefore being more effective in semen coagulum liquefaction.

Overall, our data provide support for the occurrence of an event of gene birth by duplication linked to the origin of *KLK3* in a common ancestor of Catarrhini and to an adaptive process associated to the expanded spectrum of KLK3 proteolysis. It further points to multiple events of *KLK2* death through different genomic mechanisms: Unequal crossing-over between *KLK3* and KLK2 led to the loss of *KLK2* and to the rise of a *cKLK*, whereas large deletions caused the excision of *KLK2* and relaxation of selective constraints led to *KLK2* pseudogenization. In spite of the proposed specialized role of KLK2 and KLK3 in the cascade of seminal plasma liquefaction, their substrate affinity to arginine and common patterns of expression suggest some level of redundancy for *KLK2*; however, such an argument would not explain the loss of *KLK2* observed in more ancient primate species or the skewed activity of *KLK2* in Catarrhini.

Materials and Methods

The genomic sequences from *H. sapiens, Pan troglodytes, P. paniscus, G. gorilla, Pongo pygmaeus, M. mulatta, M. fascicularis,* and *Callithrix jacchus* were retrieved from

public databases. The genomic sequences from *N. leucogenys, M. fuscata, Pap. anubis, Chlorocebus aethiops, C. guereza, Saimiri boliviensis, A. nancymaae, Cal. moloch, Ateles geoffroyi, Eulemur macaco, Lemur catta*, and *Otolemur garnettii* were obtained by Sanger-based shotgun sequencing (supplementary table S1, Supplementary Material online). BAC clones spanning the KLK2–KLK3 genomic fragment were isolated from the following libraries (see http://bacpac.chori.org, last accessed December 10, 2012), as described (Thomas et al. 2002, 2003): *P. troglodytes* (CHORI-251), *Sai. boliviensis* (CHORI-254), *Ate. geoffroyi* (UC-1), *N. leucogenys* (CHORI-271), *Cal. moloch* (LBNL-5), *C. guereza* (CHORI-272), *Pon. pygmaeus* (CHORI-253), *Pap. anubis* (RPCI-41), and *Chl. aethiops* (CHORI-252). Specifically, each library was screened using pooled sets of oligonucleotide-based probes designed from the established sequence of KLK locus. After isolation and mapping, BACs were shotgun sequenced on an ABI 3130 automated sequencer and subjected to sequence finishing, as described (Blakesley et al. 2004). KLK genes were annotated based on alignments to human RefSeq cDNA and protein sequences with the BATI algorithm (Blast, Annotate, Tune, Iterate) using four Perl scripts—Tbex, BlastSniffer, GeneTuner, and bgmix—available at http://degradome.uniovi.es/downloads.html (last accessed December 10, 2012). Briefly, BATI allows the annotation in the target genome of all orthologs and paralogs from the input set of cDNA and protein sequences. Tbex compares all the input sequences with the target genomic sequence by tblastn. BlastSniffer rebuilds each putative gene from the tblastn hits considering all the possible hit combinations and sets a raw score for each of them. GeneTuner shows the result from the previous step in the context of the template genome allowing the user to define exon/intron boundaries. Finally, bgmix creates a composite file with all the tblastn comparisons and highlights those hits overlapping defined exons. This helps the identification and annotation of novel putative genes that have not been annotated in an iterative process.

The genomic sequences spanning the *SEMG1–SEMG2* cluster were retrieved from Hurle et al. (2007) with the exception of *N. leucogenys* and *Cal. moloch* (AC198263 and AC207864, respectively), which were sequenced according to the methods described earlier for this study.

Maximum-likelihood estimates of d_N/d_S (ω) were carried out using the codeml program from the software package Phylogenetic Analysis by Maximum Likelihood—PAML version 4.2 (Yang 2007). To run PAML, we first reconstructed a phylogenetic tree using all the sequences except for *Hylobates* sp. and *L. catta* whose sequences were incomplete. To carry out a comprehensive analysis of pseudogenes, their sequences were only included after the removal of positions affected by premature stop codons and frameshift mutations. The phylogenetic tree was built using the maximum-likelihood method, implemented in DNAml, from the software package Phylogeny Inference Package (PHYLIP; http://evolution.genetics.washington.edu/phylip.html). The tree was consistent with the known primate phylogeny. To test for variable selective pressures among branches, we performed the branch model using either the null model (one ratio) or nested models (two-ratio, three-ratio, and four-ratio models) (Yang 1998; Bielawski and Yang 2003). The values of $\omega > 1$ were considered as evidences of positive selection, the values of $\omega < 1$ were regarded as an indication of purifying selection, and the values of $\omega \sim 1$ were inferred as neutral. The significance of each nested model was obtained from twice the variation of likelihoods ($-2\Delta l$) using a χ^2 statistic. To evaluate lineage-specific changes at amino acid sites, we performed the branch-site model for the anc*KLK3*. This model assumes that the branches on the phylogeny are divided a priori into foreground (anc*KLK3*) and background (remaining branches in the phylogeny) and allows ω to vary both among sites in the protein and across branches. For the branch-site model (Yang and Nielsen 2002), comparisons with critical χ^2 were carried out as described (Zhang et al. 2005). To test for variation in ω between sites of *KLK3* and functional *KLK2*, we used different codon models for each gene alone and compared neutral and selection models: M1–M2 and M7–M8 (Nielsen and Yang 1998; Yang et al. 2000). The Bayes empirical Bayes was used to calculate posterior probabilities of site classes, to identify sites under positive selection for the significant likelihood ratio tests (Yang et al. 2005). KLK3 three-dimensional model (2ZCH.pdb) was retrieved from RCSB PDB Protein Data Bank (http://www.rcsb.org/pdb/home/home.do). KLK2 three-dimensional model was generated by SwissModel (http://swissmodel.expasy.org/workspace) using KLK2 and KLK3 human sequences and KLK3 three-dimensional model (2ZCH.pdb) (Schwede et al. 2003).

Statistical analysis was performed by means of *t*-test, analysis of variance, and Spearman rank correlation.

Supplementary Material

Supplementary tables S1 and S2 and figures S1–S4 are available at *Genome Biology and Evolution* online (http://gbe.oxfordjournals.org/).

Acknowledgments

The authors thank the Lisbon Zoo for their collaboration in providing the primate samples. This work was supported by a fellowship SFRH/BD/68940/2010 from the Portuguese Foundation for Science and Technology (FCT) to P.I.M., by POPH-QREN—Promotion of Scientific Employment, by the European Social Fund, and by national funds of the Ministry of Education and Science to P.I.M. and S.S., and in part by the Intramural Research Program of the National Human Genome Research Institute. IPATIMUP (an Associate Laboratory of the Portuguese Ministry of Education and Science) is partially supported by FCT.

Literature Cited

Adzhubei IA, et al. 2010. A method and server for predicting damaging missense mutations. Nat Methods. 7:248–249.
Anderson MJ, Hessel JK, Dixson AF. 2004. Primate mating systems and the evolution of immune response. J Reprod Immunol. 61:31–38.
Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics. 3: 201–212.
Blakesley RW, et al. 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res. 14: 2235–2244.
Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. PLoS Genet. 1:e35.
Debela M, et al. 2006. Specificity profiling of seven human tissue kallikreins reveals individual subsite preferences. J Biol Chem. 281:25678–25688.
Demuth JP, Hahn MW. 2009. The life and death of gene families. Bioessays 31:29–39.
den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat. 15:7–12.
Dixson AF, Anderson MJ. 2004. Sexual behavior, reproductive physiology, and sperm competition in male mammals. Physiol Behav. 83:361–371.
Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat Genet. 36:1326–1329.
Emami N, Diamandis EP. 2007. New insights into the functional mechanisms and clinical applications of the kallikrein-related peptidase family. Mol Oncol. 1:269–287.
Harcourt AH, Harvey PH, Larson SG, Short RV. 1981. Testis weight, body weight, and breeding system in primates. Nature 293:55–57.
Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.
Hurle B, Swanson W, Program NCS, Green ED. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome Res. 17:276–286.
Janssen S, et al. 2004. Screening a combinatorial peptide library to develop a human glandular kallikrein 2-activated prodrug as targeted therapy for prostate cancer. Mol Cancer Ther. 3:1439–1450.
Jensen-Seaman MI, Li WH. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. J Mol Evol. 57: 261–270.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.
Lovgren J, Airas K, Lilja H. 1999. Enzymatic action of human glandular kallikrein 2 (hK2). Substrate specificity and regulation by Zn2+ and extracellular protease inhibitors. Eur J Biochem. 262:781–789.
Lundwall A, Brattsand M. 2008. Kallikrein-related peptidases. Cell Mol Life Sci. 65:2019–2038.
Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121–152.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.
Olsson AY, Valtonen-Andre C, Lilja H, Lundwall A. 2004. The evolution of the glandular kallikrein locus: identification of orthologs and pseudogenes in the cotton-top tamarin. Gene 343:347–355.
Pavlopoulou A, Pampalakis G, Michalopoulos I, Sotiropoulou G. 2010. Evolutionary history of tissue kallikreins. PLoS One 5:e13781.
Rawlings ND, Barrett AJ, Bateman A. 2012. MEROPS: the database of proteolytic enzymes, their substrates, and inhibitors. Nucleic Acids Res. 40:D343–D350.
Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 31: 3381–3385.
Thomas JW, et al. 2002. Parallel construction of orthologous sequence-ready clone contig maps in multiple species. Genome Res. 12: 1277–1285.
Thomas JW, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793.
Valtonen-Andre C, Olsson AY, Nayudu PL, Lundwall A. 2005. Ejaculates from the common marmoset (*Callithrix jacchus*) contain semenogelin and beta-microseminoprotein but not prostate-specific antigen. Mol Reprod Dev. 71:247–255.
Wlasiuk G, Nachman MW. 2010. Promiscuity and the rate of molecular evolution at primate immunity genes. Evolution 64:2204–2220.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.
Yousef GM, Diamandis EP. 2001. The new human tissue kallikrein gene family: structure, function, and association to disease. Endocrine Rev. 22:184–204.
Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.

Associate editor: George Zhang

Paper II - Adaptive Evolution Favoring *KLK4* Downregulation in East Asians

Mol. Biol. Evol. 2015 *Epub ahead of print*

Adaptive Evolution Favoring *KLK4* Downregulation in East Asians

Patrícia Isabel Marques,[1,2,3,4] Filipa Fonseca,[1,2] Tânia Sousa,[1,2] Paulo Santos,[1,2,] Vânia Camilo,[1,2] Zélia Ferreira,[5] Victor Quesada,[3] and Susana Seixas*[,1,2]

[1]Instituto de Investigação e Inovação em Saúde, Universidade do Porto (I3S), Porto, Portugal

[2]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal

[3]Department of Biochemistry and Molecular Biology-IUOPA, University of Oviedo, Oviedo, Spain

[4]Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal

[5]Department of Computational and Systems Biology, University of Pittsburgh

***Corresponding author:** E-mail: sseixas@ipatimup.pt.

Associate editor: Anna Di Rienzo

Abstract

The human *kallikrein* (*KLK*) cluster, located at chromosome 19q13.3–13.4, encodes 15 serine proteases, including neighboring genes (*KLK3*, *KLK2*, *KLK4*, and *KLK5*) with key roles in the cascades of semen liquefaction, tooth enamel maturation, and skin desquamation. *KLK2* and *KLK3* were previously identified as targets of adaptive evolution in primates through different mechanisms linked to reproductive biology and, in humans, genome-wide scans of positive selection captured, a yet unexplored, evidence for *KLK* neutrality departure in East Asians. We perform a detailed evaluation of *KLK3*–*KLK5* variability in the 1000 Genomes samples from East Asia, Europe, and Africa, which was sustained by our own sequencing. In East Asians, we singled out a 70-kb region surrounding *KLK4* that combined unusual low levels of diversity, high frequency variants with significant levels of population differentiation ($F_{ST} > 0.5$) and fairly homogenous haplotypes given the large local recombination rates. Among these variants, rs1654556_G, rs198968_T, and rs17800874_A stand out for their location on putative regulatory regions and predicted functional effects, namely the introduction of several microRNA binding sites and a repressor motif. Our functional assays carried out in different cellular models showed that rs198968_T and rs17800874_A operate synergistically to reduce *KLK4* expression and could be further assisted by rs1654556_G. Considering the previous findings that *KLK4* inactivation causes enamel malformations in humans and mice, and that this gene is coexpressed in epidermal layers along with several substrates involved in either cell adhesion or keratinocyte differentiation, we propose *KLK4* as another target of selection in East Asians correlated to tooth and epidermal morphological traits.

Key words: **serine proteases, human adaptation, sequence variation, *cis*-regulation, morphological traits, kallikreins.**

Introduction

Kallikreins (KLKs) belong to a large family of serine proteases (S01), which have been correlated with functions in the cascades of semen liquefaction and skin desquamation, in tooth enamel formation, in neural plasticity, and in the regulation of blood pressure. Likewise, important pathologies such as atopic dermatitis, cancer, hypertension, and neurodegenerative diseases have already been associated with temporal and spatial modifications in the activity of these proteases (Borgono et al. 2004; Emami and Diamandis 2007). In humans, the subgroup of tissue *KLKs* is located in a syntenic cluster at 19q13.3–13.4, which spans over 265 kb and includes 15 paralog genes (*KLK1*–*KLK15*), encoding inactivated trypsin- or chymotrypsin-like proteases (pro-enzymes or zymogens) and a transcribed pseudogene (*KLKP1*) (fig. 1) (Emami and Diamandis 2007; Lawrence et al. 2010).

KLK genes display a common organization and, accordingly, were proposed to have originated from a single ancestor by a series of duplication events that occurred at different moments of vertebrate evolution. Indeed, most *KLK* genes are widespread across mammalian genomes and are likely to represent ancient duplicates. Still, *KLK3* shares a high level of sequence identity to *KLK2* (79%) and it stands out as the most recent member of the cluster, having arisen in the crown of Old World monkeys (Elliott et al. 2006; Pavlopoulou et al. 2010; Marques et al. 2012). In this cluster, *KLKP1* has a complex origin, in which only two in five exons are products of partial duplications of *KLK1* (Lu et al. 2006; Kaushal et al. 2008). The timing for the emergence of *KLKP1* has not yet been fully resolved, but the identification of orthologs in several nonhuman primate genomes indicates that at least it predates human speciation (Lundwall 2013).

To date, KLKs have a remarkable functional diversity and are thought to hydrolyze a plethora of substrates including enzymes, structural proteins, hormones, and cytokines (Emami and Diamandis 2007; Lawrence et al. 2010; Fortelny et al. 2014). Nevertheless, KLKs still share some key elements fundamental to their activity as zymogens, including a conserved catalytic triad (H57, D102, and S195 residues—chymotrypsin numbering) and a short N-terminal

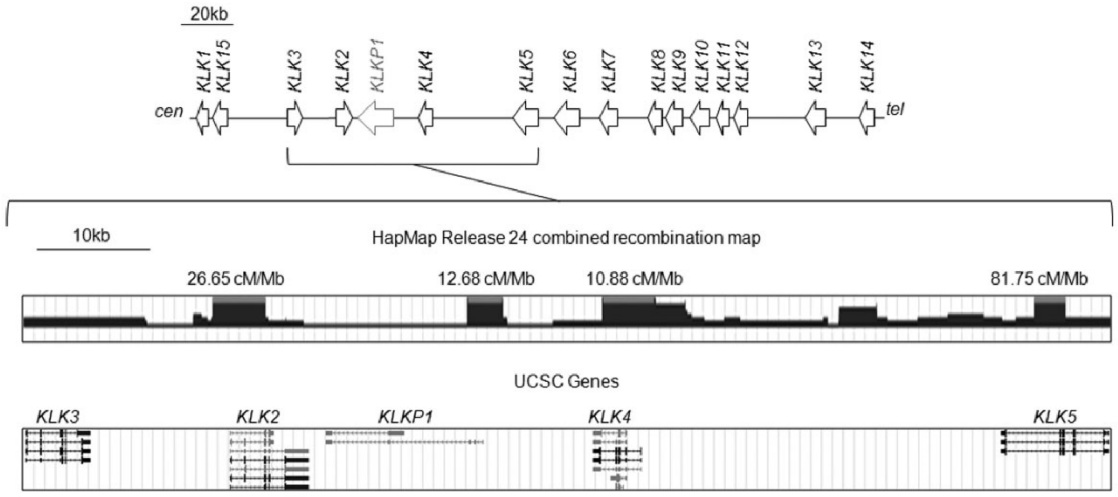


Fig. 1. Schematic representation of the human *KLK* gene cluster located at chromosome 19q13.3–13.4. Upper diagram shows the relative position of *KLK* genes. As depicted, the cluster includes 15 coding genes (black arrows) and one expressed pseudogene (gray arrow). The inset shows the *KLK3*–*KLK5* region within the UCSC Genome Browser view for recombination maps from HapMap release 24 and UCSC gene transcripts.

pro-peptide sequence. Importantly, it is this last element that allows the proteolytic regulation of KLKs, given that only after its cleavage the protease becomes active through the catalytic fissure opening (Lundwall and Brattsand 2008). The *KLKs* tissue expression profiles also show significant differentiation: *KLK2*, *KLK3* and *KLKP1* are mainly prostate-specific genes, whereas *KLK4*, *KLK5*, and the remaining genes are highly expressed in prostate and/or testis but also found in many other tissues (Lu et al. 2006; Shaw and Diamandis 2007; Kaushal et al. 2008).

The neighboring genes *KLK3*, *KLK2*, *KLK4*, and *KLK5* have been directly or indirectly implicated in male reproductive biology. Specifically, *KLK3*, *KLK2*, and *KLK5* are known to have a crucial role in the proteolytic cascade of semen liquefaction through the hydrolysis of semen coagulum proteins, mostly semenogelin 1 and 2 (SEMGs), allowing the release of spermatozoa (Lilja 1985; Deperthes et al. 1996; Michael et al. 2006). In addition, *KLK2*, *KLK4*, and *KLK5* are also thought to have a critical role in the activation of pro-KLK3 (Lovgren et al. 1997; Takayama et al. 2001; Michael et al. 2006) and the whole region from *KLK3* to *KLK5* seems to be prominently androgen-regulated (Yousef and Diamandis 1999; Shaw and Diamandis 2008; Lawrence et al. 2012).

In addition, *KLK4* and *KLK5* also participate in other important biological processes. For instance, KLK5 is recognized to have an essential activity in the skin desquamation cascade, enabling the shedding of old epidermal layers through the degradation of cell adhesion structures (Caubet et al. 2004; Eissa and Diamandis 2008). On the other hand, *KLK4* has been described as the most pervasive member of the family, found in several epidermal layers and also implicated in tooth enamel maturation (Komatsu et al. 2003; Obiezu et al. 2005; Lu et al. 2008; Fortelny et al. 2014). In the latter case, *KLK4* inactivation has been associated with a subtype of *amelogenesis imperfecta (2A1)*, which is characterized by a clinical phenotype that includes abnormal tooth development due to enamel hypomaturation, tooth brownish hyperpigmentation, and hypersensitivity to hot and cold stimuli (Hart et al. 2004; Wang et al. 2013).

Noticeably, *KLK2* and *KLK3* were previously recognized as targets of adaptive evolution in primates, due to their importance in the cascade of semen liquefaction. Specifically, several residues located in key regions of these molecules were found to be positively selected, and correlations between the loss and gain of *KLK2/3* function and the number of variable repeat units in SEMGs, or the mating system adopted by each species, were also observed (Clark and Swanson 2005; Marques et al. 2012). A possible role of *KLKs* in human adaptation is also conceivable if one considers the overrepresentation of molecular processes linked to reproduction (fertility, gametogenesis, and spermatogenesis) in genome-wide scans (GWS) of positive selection (Voight et al. 2006; Wang et al. 2006), as well as the understanding that several skin and tooth traits were shaped by selective forces during recent human history (McEvoy et al. 2006; Kimura et al. 2007; Sabeti et al. 2007; Soejima and Koda 2007; Hider et al. 2013; Gautam et al. 2015). In fact, three independent GWS of positive selection captured some features suggestive of a selective signature in the *KLK2–KLK5* region in East Asians (Voight et al. 2006; Pickrell et al. 2009; Pybus et al. 2014). Even though *KLKs* may not be among the top extreme values of empirical distributions (1%), currently, it is well accepted that the majority of GWS captured mostly classic selective sweeps, which may be "the tip of the iceberg," whereas many other unsuspected cases of natural selection remain to be identified (Pritchard et al. 2010; Hernandez et al. 2011; Messer and Petrov 2013). Interestingly, the most striking examples of positive selection specific to Asians are associated with *SLC24A5* and *EDAR* genes, and, while the first contributes to lighter skin color (Soejima and Koda 2007), the latter has been correlated to

hair and tooth morphology and sweating traits (Sabeti et al. 2007; Fujimoto, Kimura, et al. 2008; Fujimoto, Ohashi, et al. 2008; Kimura et al. 2009; Park et al. 2012; Kamberov et al. 2013). However, other examples of positive selection, possibly linked to reproductive and skin physiology, have been reported for Asian populations (Ferreira et al. 2013; Hider et al. 2013).

In this study, we sought to gain a better understanding of the evolutionary forces acting on the *KLK3–KLK5* segment of the 19q13.3–13.4 cluster, by focusing on the still unexplored signatures of positive selection identified by previous GWSs in East Asians. To this end, we combined an in-depth analysis of *KLK* genetic variability (available databases and our own sequencing) and functional assays for the most likely candidate variants of selection. Our results provide support for a nonneutral evolution of the *KLK* cluster and advance *KLK4* downregulation as an adaptive mechanism with probable significance in tooth and epidermal human traits.

Results

KLK genes were previously associated with unusual patterns of genetic diversity in East Asians, highlighting a possible signature of natural selection that was never explored before. Initially, *KLK2* and *KLK4* were found among the top 5% counts of an empirical GWS for recent positive selection ($P = 0.047634$ for *KLK2* and *KLK4*), based on the integrated haplotype score (iHS) statistic for the CHB+JPT sample (CHB: Han Chinese in Beijing, China; and JPT: Japanese in Tokyo, Japan) in HapMap phase II (http://haplotter.uchicago.edu/, last accessed December 19, 2013) (Voight et al. 2006). Later on, a region spanning from *KLK2* to *KLK5* was found to display relatively high cross-population extended haplotype homozygosity (XP-EHH) scores in East Asians (supplementary fig. S1*A*, Supplementary Material online) in a study based on the data from approximately 600,000 single nucleotide polymorphisms (SNPs) genotyped for the 52 populations from the Human Genome Diversity Panel (Li et al. 2008; Pickrell et al. 2009). In a more recent survey, already using the data from the 1000 Genomes Project (1000G) phase I, for three main populations (CHB; YRI: Yoruba in Ibadan, Nigeria; and CEU: Utah residents with ancestry from northern and western Europe; "The 1000 Genomes Selection Browser 1.0"; http://hsb.upf.edu/, last accessed April 10, 2014), the same region encompassing *KLK2–KLK5* genes showed several footprints of selection in the CHB sample (1000 Genomes Project Consortium et al. 2012; Pybus et al. 2014). These included significantly high empirical rank scores for F_{ST}, iHS, XP-EHH, and cross-population composite likelihood ratio statistics (rank score ≥ 2) (supplementary fig. S1*B* and table S1, Supplementary Material online).

Patterns of Genetic Variation at *KLK3–KLK5* Genes

To better explore the hallmarks of natural selection identified in previous studies, we surveyed the *KLK* genetic variation (*KLK3–KLK5*: chr19:51353000–51461000, GRCh37/hg19) available at the 1000G phase I catalog (1000 Genomes Project Consortium et al. 2012), and we centered our analysis in the three populations better characterized from a demographic point a view, and representing the major human groups, Africans (YRI), Europeans (CEU), and Asians (ANS:CHB+JPT). Overall for the *KLK3–KLK5* region, the 1000G data comprised a total of 1,419 SNPs, including a nonsense mutation and 25 nonsynonymous substitutions, in which five were low-frequency variants ($f \leq 0.015$) predicted as deleterious by SIFT and Polyphen (Kumar et al. 2009; Adzhubei et al. 2010) (supplementary table S2, Supplementary Material online). Worth to note, the nonsense mutation was located in *KLK4* (rs104894704; NP_004908.4: p.W153*) and reported as associated with *amelogenesis imperfecta* (Hart et al. 2004).

In our approach, to evaluate possible deviations from neutral expectations of the site frequency spectrum (SFS), we calculated summary statistics of polymorphism levels per gene and per population (table 1 and supplementary table S3, Supplementary Material online, for other 1000G surveyed populations). As it would be expected from the "out-of-Africa" model of human demography, higher levels of sequence variation (θ_W and π) were found in Africa. Nonetheless, several statistics for the ASN population were found to depart from neutral expectations, as assessed by coalescent simulated null-distributions for constant population size (conservative model), and for two best-fit models of East Asian demography, in which CHB and JPT were merged into a single sample (Laval et al. 2010; Gravel et al. 2011). Specifically, for *KLKP1* in the ASN population, the Tajima's *D* statistic (Tajima 1989) and Fay and Wu's *H* test (Fay and Wu 2000; Zeng et al. 2006) were associated with strongly negative values ($D_{KLKP1} = -1.67$; $H_{KLKP1} = -3.14$), indicating that *KLKP1* combines an excess of rare variants with high-frequency derived alleles, and which may be interpreted as an evidence of nonneutral evolution. Other genes found to depart from neutral expectations in the ASN sample were *KLK3* and *KLK5* ($D_{KLK3} = 1.66$ and $H_{KLK5} = -4.02$, respectively).

To control for a possible ascertainment bias of the 1000G data, introduced by the inclusion of segments with low sequence coverage, as well as to evaluate the potential effects in variant calling of the usage of next-generation methods in the screening of regions with stretches of considerable homology, we Sanger-sequenced 19 segments encompassing the *KLK3–KLK5* region (~16.5 kb per individual), in a subsample of YRI, CEU, and ASN individuals (supplementary table S4, Supplementary Material online). Only minor discrepancies were observed between data sets where the 1000G data failed to report ten low-frequency variants, two SNPs at intermediate frequencies, and a copy number variation (CNV) previously described in the 3′-UTR of *KLK4* (Levy et al. 2007) (supplementary table S5, Supplementary Material online).

The statistics based on the SFS, which are more likely to be sensitive to the 1000G sequencing scheme, were found to provide a good overlap between the two sequencing approaches for the estimators of the population mutation rate parameter ($\theta = 4N_e\mu$) (supplementary table S6, Supplementary Material online). The neutrality tests, in general, were shown to display similar trends (negative or positive values) still, the concordance in the statistical significance, as addressed by coalescent simulated null distributions for

Table 1. Summary Statistics of *KLK3–KLK5* Population Variation from 1000G Data.

	ASN							CEU							YRI						
	N	*S*	θ_w	π	*D*	*D**	*H*	*N*	*S*	θ_w	π	*D*	*D**	*H*	*N*	*S*	θ_w	π	*D*	*D**	*H*
KLK3 (5,849 bp)	372	48	12.64	20.08	1.66*	−0.83	−0.24	170	54	16.17	21.33	0.97	−0.37	−0.55	176	68	20.24	17.36	−0.44	−0.23	−0.62
KLK2 (7,319 bp)		46	9.68	8.07	−0.47	−1.61	−0.60		40	9.57	11.82	0.70	−0.29	0.33		63	14.98	14.09	−0.18	0.45	0.55
KLKP1 (14,303 bp)		75	8.07	3.44	−1.67*	−0.80	−3.14*		64	7.84	10.57	1.07	0.49	−0.87		137	16.67	18.49	0.34	0.92	0.35
KLK4 (4,387 bp)		29	10.18	10.56	0.10	−0.68	0.08		27	10.78	16.11	1.42	0.71	0.31		37	14.68	18.47	0.76	0.49	−0.08
KLK5 (9,786 bp)		54	8.50	10.10	0.54	−0.49	−4.02†*		55	9.84	12.43	0.80	−0.80	−2.82†*		89	15.83	15.32	−0.10	0.43	−0.86

NOTE.—*N*, number of chromosomes; *S*, number of segregating sites; θ_w, Watterson's estimator of θ (Watterson 1975) per base pair ($\times 10^{-4}$); π, nucleotide diversity per base pair ($\times 10^{-4}$); *D*, Tajima's *D* statistic (Tajima 1989); *D**, Fu and Li *D** statistic (Fu 1997). *H*, Fay and Wu's *H* test (Fay and Wu 2000; Zeng et al. 2006). *P* values < 0.05 according to the constant size model are underlined.

†Significant *P* values < 0.05 according to Gravel model (Gravel et al. 2011) with recombination; *Significant *P* values < 0.05 according to Laval model (Laval et al. 2010) with recombination.

constant population size and East Asian demography, was not complete (supplementary table S6, Supplementary Material online). As example, *KLKP1*, which presented the most robust results, also surpassing *H* test in the Sanger-sequencing data set ($H_{KLKP1} = -5.36$), lost Tajima's *D* significance to a less strong negative value. On the other hand, *KLK4*, the gene found to harbor the highest number of failed variants in the 1000G (supplementary table S5, Supplementary Material online), changed the Tajima's *D* score in the Sanger-sequencing data set to a slight negative value, almost reaching significance in the Fu and Li *D** statistic ($P = 0.05$; Laval et al. 2010, best-fit model). Under this juncture, it is important to stress out that some inconsistencies may be partly explained by the larger number of individuals analyzed in the 1000G, as well as, the larger regions included in 1000G gene-centered analysis in comparison to our Sanger-sequencing screening. Even so, collected data seem to fit previous findings from GWS of positive selection and suggest a trend toward low-frequency variants and/or an excess of high-frequency-derived alleles across *KLK2*, *KLKP1*, *KLK4*, and *KLK5* sequences in East Asians (table 1, supplementary tables S3 and S6, Supplementary Material online). We must emphasize that even though Fay and Wu *H* and Tajima's *D* neutrality tests show increased power to detect positive selection signatures, as long as the advantageous variant is swept to higher frequencies, Fay and Wu *H* test is usually the most powerful statistic before the selected variants reach fixation ($f > 0.6$) (Fay and Wu 2000; Przeworski 2002; Zeng et al. 2006).

Tests for the Signature of Natural Selection

To obtain a more comprehensive analysis of the genetic variation in the region spanning from *KLK3* to *KLK5* genes, we calculated the nucleotide diversity (π) (fig. 2*A*) and Tajima's *D* (fig. 2*B*) statistics in a sliding window approach. In general, the entire region from *KLK2* to *KLK5* displays reduced diversity levels and negative Tajima's *D* values for the ASN population, with *KLKP1* showing the lowest values, in comparison with CEU and YRI populations. Taking into account that natural selection may be acting on the ASN population alone and a hypothetical beneficial variant may have been swept to higher frequencies in ASN but not in YRI or CEU, we next examined the levels of population differentiation as measured by F_{ST} statistic (ASN vs. non-ASN populations in fig. 3*A*; ASN vs. CEU, ASN vs. YRI, and CEU vs. YRI in supplementary fig. S2, Supplementary Material online). While using the ASN versus non-ASN (CEU+YRI) comparison as a preferred F_{ST} statistic, we assigned CEU and YRI samples into a single group and ASN sample into another group. This method of clustering different populations has been previously described (Excoffier et al. 2009) and uses an explicit hierarchical island model, found to be robust to heterogonous affinities between populations and variable migrations rates within, and between groups. In addition, even when an apparently unrealistic population structure is assumed, this method has a lower false positive discovery rate than a simple island model, and it has been recognized as a useful method in the identification of populations and/or SNPs contributing to global F_{ST} values (Excoffier et al. 2009). Therefore, by performing the ASN versus non-ASN comparison, we are more likely capturing those SNPs representing true East Asians adaptations, than by carrying out two-by-two population analysis.

Surprisingly, the most extreme and significant F_{ST} values ($F_{ST} > 0.50$ with *P* values < 0.05) in the ASN versus non-ASN comparison were not observed in *KLKP1* but in two other regions: The first one, close to *KLK4* (chr19:51404000-51411000), and the second one within an intergenic region downstream of *KLK5* (chr19:51434000–51450000) (fig. 3*A*). Notably, all these F_{ST} scores are superior to the previously described averages for ASN and non-ASN comparisons (0.08–0.15) (1000 Genomes Project Consortium et al. 2012) and overlap with the highest empirical ranks scores obtained for the 1000G global comparisons (fig. 3*B* and supplementary table S1, Supplementary Material online) (Pybus et al. 2014). Therefore, the hypothesis of the *KLK3–KLK5* sublocus as a potential target of selection in the ASN population seems to be reinforced by F_{ST} analysis.

Even though we could not detect an obvious candidate variant, we excluded a priori two regions, one encompassing *KLK3* and partially *KLK2* (exon I to exon III) and the other one comprising *KLK5* (exon I to intron V) due to the presence of two recombination hotspots (26.65 and 81.75 cM/Mb; fig. 1) setting the limits of an approximately 70-kb region (chr19:51378273–51451045) with multiple significant SFS and F_{ST} statistics. Two other hotspots were identified near *KLK4* but these are associated with lower recombination rates (12.68 and 10.88 cM/Mb; fig.1), and somewhat less strong

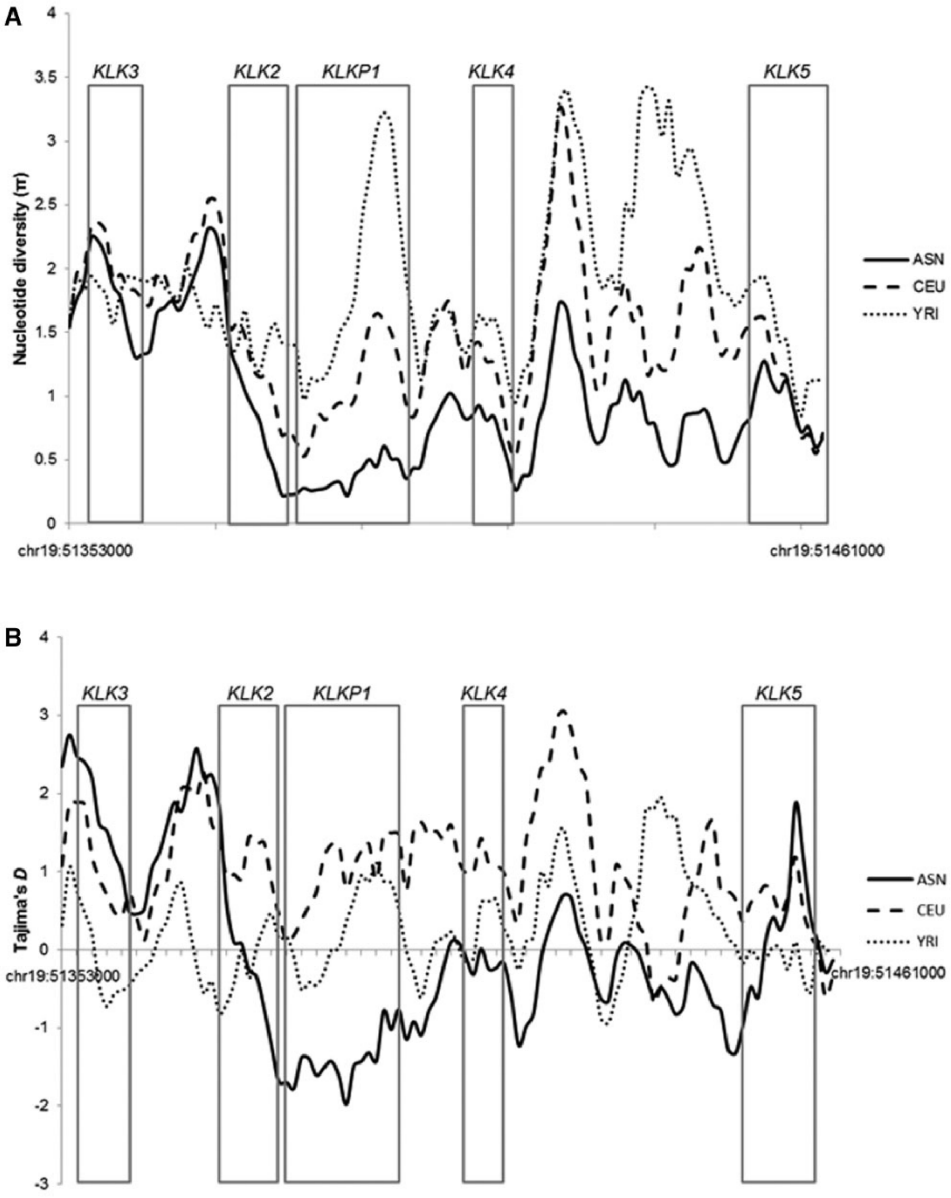


Fig. 2. Sliding window of nucleotide diversity per base pair ($\times 10^{-3}$) (*A*) and Tajima's *D* (*B*) in the *KLK3*–*KLK5* region in ASN (CHB+JPT), CEU, and YRI (solid, dashed, and dotted lines, respectively). Window size: 5,000 bp; increment: 1,000 bp.

effects in linkage disequilibrium (LD) structure (supplementary fig. S3, Supplementary Material online).

In a different approach, we calculated standardized iHS statistics using the data for the whole chromosome 19. After normalization, 25 of 204 SNPs located in the 70-kb region presented high iHS values ($|\text{iHS}| > 2$; supplementary table S7, Supplementary Material online). As selective sweeps tend to produce clusters of significant iHS scores across target regions, and selected sites not necessarily show top extreme or even significant values, we decided to divide chromosome 19 into nonoverlapping windows of 100 and 50 kb and calculate the proportion of SNPs with $|\text{iHS}| > 2$, as previously described by others (Voight et al. 2006; Kudaravalli et al. 2009; Szpiech and Hernandez 2014). The windows of 100 kb (chr19:51400001–51500001) and 50 kb (chr19:51400001–51450001) containing potential target variants were found to be among the top 20% and 10% of the empirical distributions, respectively. The 100-kb window ranges from *KLK4* to *KLK8*, spans the 81.75 cM/Mb recombination hotspot, and thus it contains another *KLK* segment with an independent evolutionary history (*KLK6*–*KLK8*). Conversely, the 50-kb window is mainly defined by the hypothetical selected

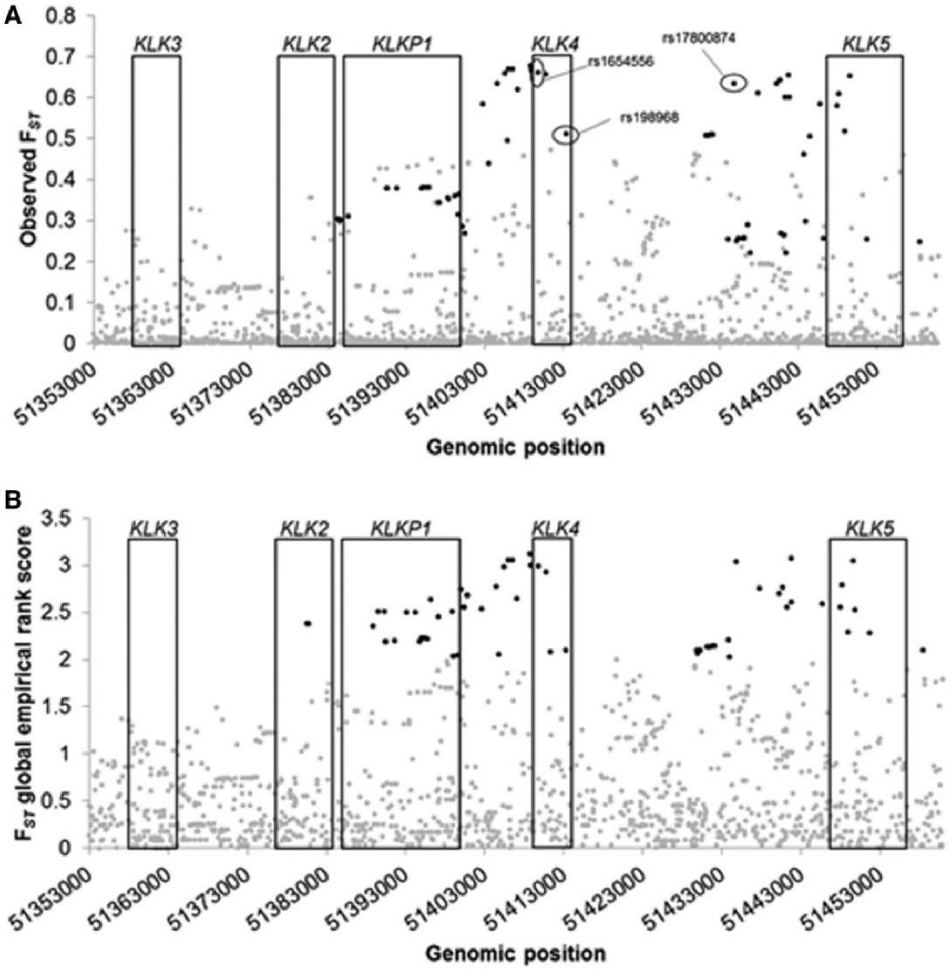


FIG. 3. Genetic population differentiation (F_{ST}) analysis for *KLK3–KLK5* locus of ASN versus non-ASN populations (*A*) and empirical rank F_{ST} scores based on global comparisons for CHB, CEU, and YRI (*B*). Genes location are delimited by open boxes. SNPs with significant F_{ST} *P* values (upper $P < 0.05$) or significant empirical rank scores (http://hsb.upf.edu/) are displayed in black.

region encompassing *KLK4* gene. Hence, the lower rank obtained for the 100-kb window (90/590) in comparison to the 50-kb window (115/1,180) most probably results from the inclusion of a nonselected region, where there is a lower proportion of SNPs with $|\text{iHS}| > 2$.

Importantly, the occurrence of four recombination hotspots in a short segment of approximately 70 kb is likely to have prevented the inference of selective signatures by an LD-based statistic, such as iHS, and empirical comparisons using chromosome 19 data. Indeed, other authors have already demonstrated that long-range haplotype (LRH) tests and GWS studies tend to discover selected loci in regions of low recombination, lacking robustness to account for the large variability in local recombination rates (Carlson et al. 2005; Kelley et al. 2006; O'Reilly et al. 2008; Ferrer-Admetlla et al. 2014). To overcome this limitation, we performed the DIND (Derived Intra-allelic Nucleotide Diversity) test for the 70-kb region, controlled by coalescent simulated data under constant population size and best-fit models for East Asian demography (Laval et al. 2010; Gravel et al. 2011), and assuming the high local recombination rates, as inferred by human fine-scale maps (McVean et al. 2004). Briefly, the DIND test calculates, per SNP, the ratio of the ancestral to derived allele nucleotide diversity ($i\pi A/i\pi D$) and has the advantage over other selection statistics of accessing for each allele its linked SFS, building-on other neutrality tests with low power to detected partial selective sweeps (Barreiro et al. 2009). Importantly, the DIND test also allows the identification of potential candidate alleles without a prior hypothesis and has been recently shown to be robust to next-generation sequence coverage variation and insensitive to low coverage (Barreiro et al. 2009; Fagny et al. 2014). In our case, the DIND test placed multiple SNPs with high derived allele frequencies (DAF > 0.70) as probable targets of selection,

providing statistical support for lower linked variation than expected under neutrality within the 70-kb region (fig. 4 and supplementary table S8, Supplementary Material online).

Next, in the absence of any common nonsynonymous variant that could have been driven to higher frequencies by natural selection, we directed our attention to regulatory regions and possible expression quantitative trait loci. Therefore, to identify the best selection candidate variants we merged the collected information for SNPs with top F_{ST}, iHS and DIND scores, with previous evidences from GWS of positive selection (iHS scores from Hapmap phase II and F_{ST} statistic from "The 1000 Genomes Selection Browser 1.0"), and ENCODE data regarding chromatin segmentation, which integrates ChIP-seq data for nine factors using a Hidden Markov Model. Specifically, we detected 31 SNPs with significant F_{ST} scores for the ASN versus non-ASN comparison ($F_{ST} > 0.50$; P values < 0.05) and for the empirical Global F_{ST} rank score, in which 27 were shown to also have significant iHS and/or DIND values (figs. 3 and 4 and supplementary table S9, Supplementary Material online). Among the high frequency-derived alleles only rs198968 and rs17800874 combined two other significant statistics beside F_{ST} scores and among common ancestral alleles with significant iHS values, the rs1654556 was the most interesting candidate according to ENCODE data (supplementary table S9, Supplementary Material online).

The variants rs198968 and rs1654556 are both located in *KLK4*, in intron I and 3′-UTR, respectively, and although the first lies in a weak promoter (*H1-hESC*: H1 human embryonic stem cells), the latter is located in an insulator shared by all cell lines analyzed by ENCODE (supplementary fig. S4 and table S9, Supplementary Material online). Moreover, rs198968 and rs1654556 variants have already been predicted to alter the binding site of a transcription factor (BTB/POZ domain) and miRNAs (miR-1254, 378, 422a, 661), respectively (Lose et al. 2012).

The remaining variant, rs17800874, was found to be located in the intergenic region between *KLK4* and *KLK5* within a putative enhancer as indicated by ENCODE (NHEK: normal human epidermal keratinocyte; and HMEC: human mammary epithelial cells) (supplementary fig. S4 and table S9, Supplementary Material online). This variant had no associated binding prediction but it has been previously associated with a significant iHS value (−2.05) in HapMap phase II (Voight et al. 2006) and a high CMS score (composite of multiple signals, 8.12), which incorporates at once five selective statistics calculated for 1000G data (http://www.broadinstitute.org/mpg/cmsviewer/, last accessed June 5, 2015) (Grossman et al. 2010).

As it would be expected from the F_{ST} scores, candidate variants rs1654556_G, rs198968_T, and rs17800874_A all presented high frequencies in the East Asian samples in contrast to their reduced frequencies in worldwide populations (fig. 5A and supplementary fig. S5, Supplementary Material online). However, the LD levels for the three SNPs in the ASN sample were not impressive (rs1654556–rs198968: $D' = 0.76$, $r^2 = 0.43$; rs1654556–rs17800874: $D' = 0.28$, $r^2 = 0.07$; and rs198968–rs17800874: $D' = 0.42$; $r^2 = 0.13$). Still,

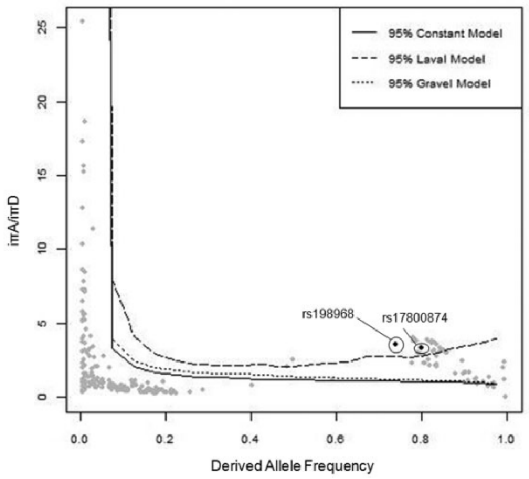


FIG. 4. Ratio of intra-allelic diversity associated with the ancestral and derived alleles (iπA/iπD) plotted as a function of the DAF in the ASN (CHB+JPT) population. Black points: Candidate SNPs rs198968 and rs17800874. $P < 0.05$; solid line: 95% constant model; dashed line: 95% Laval mode (Laval et al. 2010); dotted line: 95% Gravel model (Gravel et al. 2011).

the haplotype configurations rs1654556_G–rs198968_T–rs17800874_A (GTA) and rs198968_T–rs17800874_A (GTA and ATA) represent in the ASN population 228/372 chromosomes (61%) and 243/372 chromosomes (65%), respectively (fig. 5B). In the CEU, these configurations reached 4% and 6% and were nearly absent in the YRI. Noteworthy, the limited variation observed within GTA and ATA haplotype configurations can be reconciled with the previous findings from the SFS (trend to negative neutrality tests) and probably suggests a partial selective sweep hypothesis.

Finally, to confirm GTA/ATA LRH structure, we calculated the EHH/REHH statistics (Sabeti et al. 2002) for an approximated 0.1-cM genetic distance around the core haplotypes, with cores defined by at least three SNPs and no more than ten. For the three candidate SNPs, a slower decay of EHH from the common core haplotype (73–78%) was observed (supplementary fig. S6, Supplementary Material online), together with a remarkable haplotype extension (chr19:51378986–51441046 [~62 kb], 51392135–51451045 [~59 kb], and chr19:51378273–51441046 [~63 kb]), considering the complexity of the *KLK3–KLK5* region. Even so, all three cores were found not to reach statistical significance in the empirical comparisons with chromosome 19 data (P values range: 0.068–0.338). Nevertheless, given that LD-based statistics tend to lack power in regions with large and variable recombination rates, these results are not entirely unexpected (Carlson et al. 2005; Kelley et al. 2006; O'Reilly et al. 2008; Ferrer-Admetlla et al. 2014).

Functional Effects of the Candidate Variants

To assess the functional impact of the candidate variants rs198968, rs17800874, and rs1654556, we preformed luciferase reporter assays in prostate (LNCaP), cervix (HeLa), and gastric (AGS) cell lines using different haplotype configurations

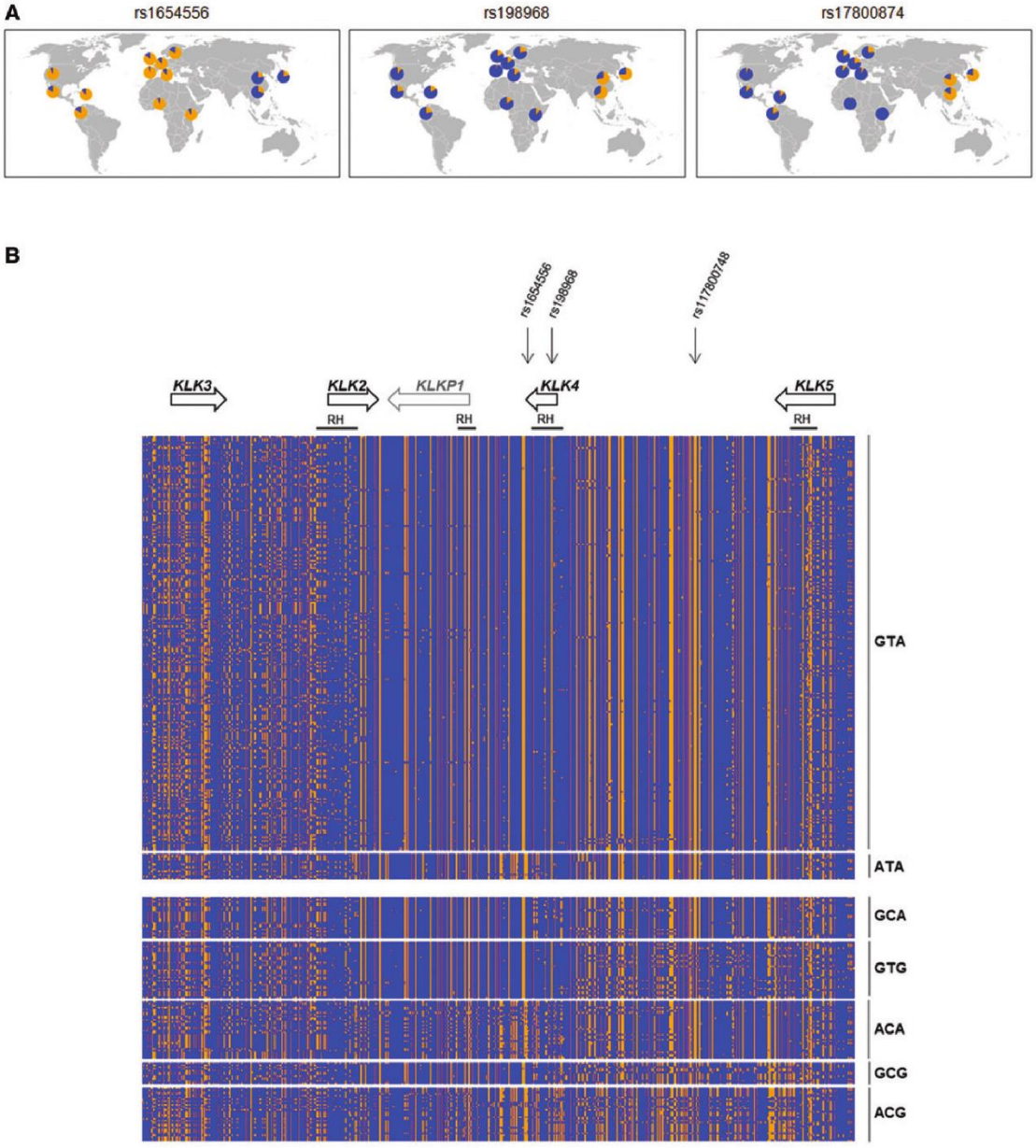


Fig. 5. Signatures of natural selection at *KLK3–KLK5* locus in human populations. (*A*) Worldwide estimated allele frequencies from 1000G data for variants rs1654556, rs198968, and rs17800874 in 14 human populations (Asia: CHB, JPT, and CHS—Southern Han Chinese in China; Africa: YRI and LWK—Luhya in Webuye, Kenya; Europe: CEU, GBR—British in England and Scotland, FIN—Finnish in Finland, IBS—Iberian populations in Spain, TSI—Toscani in Italia; Americas: ASW—African Ancestry in Southwest United States of America; CLM—Colombian in Medellin, Colombia, MXL—Mexican ancestry in Los Angeles, CA, PUR—Puerto Rican in Puerto Rico). (*B*) Schematic representation of ASN (CHB+JPT) haplotypes for *KLK3–KLK5* region. Each line represents a haplotype and columns indicate polymorphic positions. Haplotypes are organized by different configurations of rs1654556, rs198968, and rs17800874 alleles. The relative positions of *KLK* genes are depicted by open arrows, the candidate SNPs by the filled arrows and the recombination hotspots (RH) are also shown. Ancestral alleles are represented in blue and derived alleles in orange.

(fig. 6*A*). Briefly, for rs198968 constructs we cloned a genomic region of approximately 1 kb encompassing part of intron I and the beginning of exon II (chr19:51412559–51413597), and for rs17800874 we amplified approximately 1 kb of an expected enhancer region (chr19:51434619–51435702). Additionally, we analyzed the joint effect of these two SNPs by subcloning the rs198968 inserts into to the 3′-end of the rs17800874 segment, generating constructs with either derived or ancestral alleles (fig. 6*A*). For rs1654556, we generated three constructs of different lengths (~1 kb) due to the close

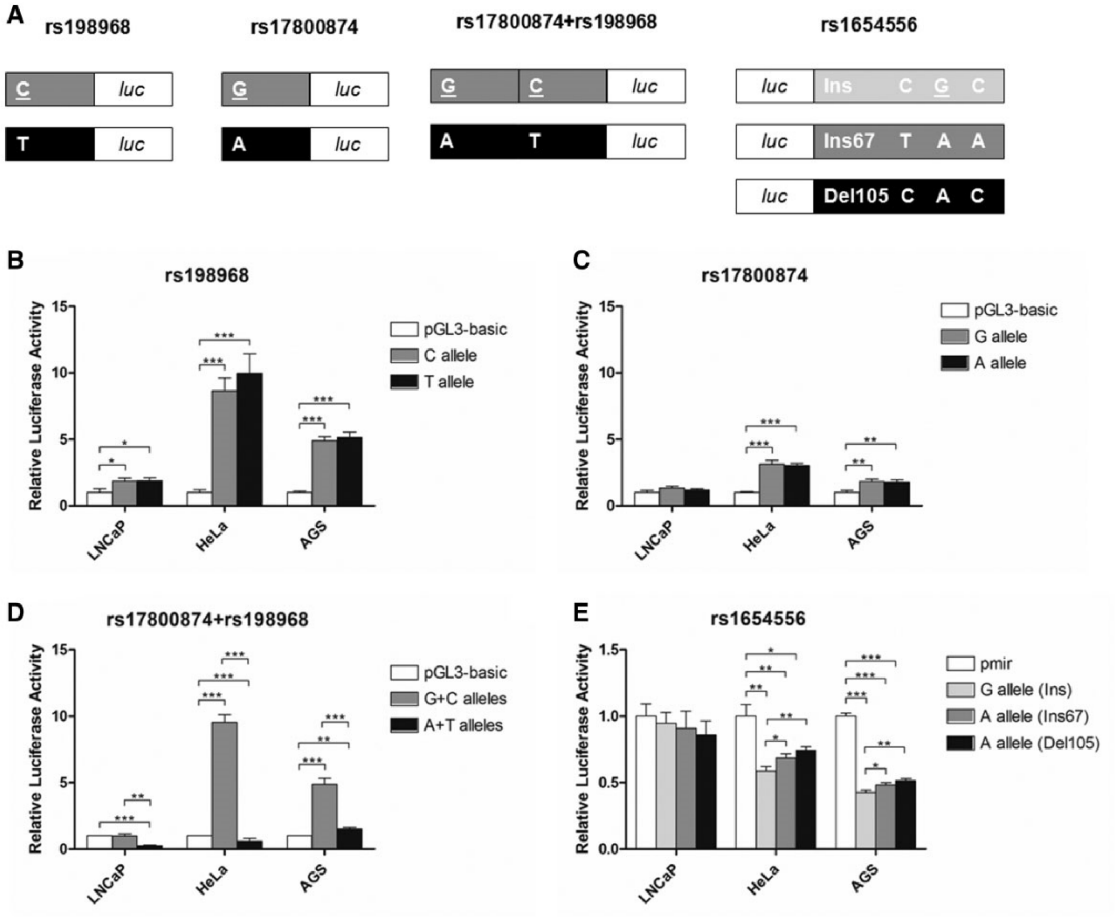


Fig. 6. In vitro validation of candidate variants by luciferase reporter assays. (*A*) pGL3 and pmirGLO constructs containing the ancestral (underlined) or the derived allele. The CNV alleles 105-bp deletion (Del105) and 67-bp insertion (Ins67) included in pmirGLO constructs are shown. Relative luciferase activity of variants rs198968 (*B*), rs17800874 (*C*), rs17800874 + rs198968 (*D*), and rs1654556 (*E*) in LNCaP, HeLa, and AGS cell lines. Data are expressed as the mean ± standard error mean for at least three experiments. $^{*}P < 0.05$; $^{**}P < 0.01$; and $^{***}P < 0.0001$.

proximity to the CNV identified in the 3′-UTR of *KLK4* (174 bp upstream of rs1654556). According to our CNV screening, the candidate variant rs1654556_G was always linked to the reference sequence allele (Ins), in opposition to the rs1654556_A variant that could be associated with three different CNV alleles: A 67-bp insertion (Ins67), a 67-bp deletion (Del67), or a 105-bp deletion (Del105) (supplementary table S10, Supplementary Material online). Our constructs took into account linked variability associated with rs1654556 and the assessed extreme CNV configurations (Ins67 and Del105; fig. 6*A*).

In our experiments, the constructs containing rs198968 or rs17800874 were found to display considerable promoter activity in HeLa and AGS cells, whereas in the LNCaP cell line we detected only a modest promoter activity for the rs198968 constructs. However, no differences were observed in the expression of opposite allele variants for rs198968 or rs17800874 (fig. 6*B* and *C*). Remarkably, when rs17800874_A was associated with rs198968_T, we observed a significant decrease in luciferase expression in comparison to the opposite haplotype configuration (0.95-, 6.46- and 3.36-fold change in LNCaP, HeLa and AGS, respectively; fig. 6*D*). In addition, the 3′-UTR region spanning rs1654556 and the CNV variants showed a significant decrease in luciferase activity in HeLa (0.59- to 0.74-fold change) and AGS (0.43- to 0.51-fold change) but not in LNCaP cells (fig. 6*E*). Here, the luciferase activity associated with rs1654556_G was found to be significantly lower than that of the rs1654556_A constructs. No significant differences were observed between rs1654556_A (Ins67) and rs1654556_A (Del105) (fig. 6*E*).

To assess the impact of rs198968 or rs17800874 variants in the expression of *KLK4* in vivo, we took advantage of the recent Genotype-Tissue Expression project (GTEx), which provides a comprehensive atlas of gene expression and regulation across multiple human tissues (GTEx-Consortium 2015). In prostate samples ($N = 36$), we could verify that carriers of rs198968- or rs17800874-derived alleles (heterozygous TC and GA, respectively) showed a trend toward reduced

KLK4 expression (supplementary fig. S7, Supplementary Material online). Even though we could not use GTEx expression data to reproduce the combined effect of rs198968_T and rs17800874_A, these results seem to reinforce our findings of a candidate variant effect in the regulation of *KLK4* expression.

Altogether these findings suggest *KLK4* downregulation might have been selectively advantageous in ASN, however, this regulatory effect may only occur when rs198968_T and rs17800874_A cosegregate in the same chromosome and where the presence of rs1654556_G may contribute through a minor reduction in gene expression.

Tissue Expression Pattern of *KLK3–KLK5* Locus

To investigate *KLK2*, *KLK3*, *KLK4*, *KLK5*, and *KLKP1* expression, we screened a panel of 21 human tissues (cDNA) by reverse transcription polymerase chain reaction (RT-PCR) (supplementary fig. S8, Supplementary Material online). For *KLK2* and *KLK3*, as expected, we observed high expression in prostate and lower levels for *KLK2* in thyroid and for *KLK3* in bladder, kidney, testis and thyroid (supplementary fig. S8*A* and *B*, Supplementary Material online). Regarding *KLK4*, we could confirm its expression at variable levels in most analyzed tissues, with prostate, thyroid, intestine, uterus, testis, thymus and trachea showing high to moderate expression (supplementary fig. S8*C*, Supplementary Material online). In the case of *KLK5*, it was absent in colon, liver, adiposities, thyroid and placenta, and with uterus, esophagus and testis presenting the highest levels of expression (supplementary fig. S8*D*, Supplementary Material online). Finally, *KLKP1* displayed only modest levels of expression in uterus, prostate, and testis (supplementary fig. S8*E*, Supplementary Material online). In the first two tissues we could associate *KLKP1* expression with the longer transcript (*KLK31P*), whereas in the testis the expression was related to the shorter mRNA (*KRIP*) (Lu et al. 2006; Kaushal et al. 2008).

Discussion

The availability of large catalogs of genetic variation has allowed researchers to gain insights on how natural selection has shaped the human genome. However, previous efforts to identify selective targets have only found limited evidence of adaptive evolution in genes linked to specific morphological traits. In this study, we performed an in-depth analysis of the natural history of the *KLK3–KLK5* cluster segment, starting by a reevaluation of the most updated human genetic variation data followed by a compilation of further evidence for a departure from neutral expectations in East Asians up to the identification of a common haplotype configuration with dramatic effects in the regulation of *KLK4* expression. Here, we hypothesized that this haplotype was driven to higher frequencies in East Asians (partial selective sweep) through a possible contribution into previously described adaptive traits, namely tooth morphology and/or diverse epidermis attributes.

First, we identified several statistical arguments (e.g., SFS tests and F_{ST}) for a nonneutral evolution of the *KLK* cluster in East Asians, most of them concentrated in a relative short segment (~70 kb) encompassing multiple recombination hotspots. Still, the transition from the discovery of genomic footprints of human adaptation into the identification of molecular targets of natural selection remains a major challenge in the field of evolutionary genetics. Considering that *KLKP1* is expressed at low levels, and most likely as a noncoding RNA, we redirected our attention into other regions with stronger functional relevance. Here, rs1654556_G, rs198968_T, and rs17800874_A variants, found to co-segregate in a common and rather homogenous haplotype, stood out as possible candidate variants.

A selective hypothesis is strengthened by our in vitro validation showing that all three candidate variants can contribute to *KLK4* downregulation. Interestingly, our results suggest that such phenomenon may happen synergistically when rs198968_T and rs17800874_A are found in the same chromosome, given that, separately, both variants have no allele-specific effect in promoter activity. In our experiments, we studied the most extreme haplotype configurations, which represent the ancestral condition (C-G: YRI: 81%, CEU: 65%, and ASN: 11%) and the putatively selected haplotype in East Asians (T-A: YRI: 1%, CEU: 5%, and ASN: 65%). The other configurations show only limited variation among populations (C-A: YRI: 1%, CEU: 16%, and ASN: 14%; or T-G: YRI: 15%, CEU: 9%, and ASN: 8%), which seems to favor the rs198968_T–rs17800874_A configuration as the main target of selection. Worth of note, rs198968 and rs17800874 are located in predicted promoter and enhancer regions, respectively, and in both cases these are in close proximity to typical features of chromatin looping and insulator regulatory elements, which includes signals of CTCF and nucleosome binding, nuclear lamina interactions, and H3K27A and H3K4me1 histone makers (Yang and Corces 2012; Holwerda and de Laat 2013). Consistently, rs198968_T has already been predicted to introduce a motif for BTB/POZ (Lose et al. 2012), a binding factor domain required for the formation of DNA-looped structures between different regulatory elements in the human β-globin cluster (Yoshida et al. 1999). Although rs17800874_A has no similar prediction, our findings show that both alleles are likely to cooperate in insulator activity and long-distance interactions of regulatory elements with impact on *KLK4* expression.

A selective advantage associated with reproductive traits was first considered given the reported significance of *KLKs* in the evolution of semen coagulation rates in primates. However, in humans, there is lack of supporting evidence for the development of specializations driven by sperm competition, and the trend toward smaller testis sizes observed in Asia (Dixson 2009) is hardly correlated to *KLK4* downregulation. On the other hand, a selective advantage correlated to tooth morphology may be advanced if the enamel defects observed in *amelogenesis imperfecta* and in *Klk4* knockout mice are taken into account. In both systems, the main phenotype is the enamel softening without thickness reduction, which in mice is progressively less mineralized from the surface to the enamel–dentin junction (Hart et al. 2004; Simmer et al. 2009; Wang et al. 2013). In this regard, human

populations and East Asians, in particular, display common dental variations, such as upper central incisor shoveling, enamel extensions of the first maxillary molar, and other metric and nonmetric measures (Turner 1990; Hanihara and Ishida 2005; Hanihara 2008; Park et al. 2012).

So far, polymorphisms in *EDAR*, *ENAM* and *WNT10A* have been to some extent correlated to dental traits and like for *KLK4*, dysfunctional *ENAM*, *EDAR* and *WNT10A* cause abnormal tooth development (Kelley and Swanson 2008; Kimura et al. 2009, 2015). In addition, *ENAM* and *EDAR* were also found to display hallmarks of positive selection in non-Africans and East Asians, respectively, but for *WNT10A* an adaptive evolution model in humans may not hold true (Sabeti et al. 2007; Kelley and Swanson 2008; Kimura et al. 2009, 2015; Kamberov et al. 2013). Notably, *ENAM* encodes *enamelin*, a protein cleaved by *KLK4*, proposed to underlie an adaptive response to changes in diet through a process of enamel thinning (Kelley and Swanson 2008). Conversely, *EDAR*, which regulates the development of organs of ectoderm origin (teeth, hair, nails, and glands), has no known interaction with *KLK4*. Moreover, it is possible that dental traits associated with *EDAR* are a byproduct of the action of selection in hair structure and/or sweating glands (Kimura et al. 2009; Kamberov et al. 2013). At this point, it is important to note that the outer enamel surface is considered as a more direct target of selection by dental functions (feeding and occlusion) than the enamel–dentine junction, which is thought to be evolutionary constrained (Kraus 1952; Smith et al. 1997; Olejniczak et al. 2007; Morita et al. 2014). In addition, the process of enamel formation has already been proposed to play a significant role in tooth shape and structure, to vary in rate and time period, and to have a nonconstant influence among teeth (Morita et al. 2014). Hence, in view of *KLK4* importance in the enamel maturation stage, which allows crystallites to grow in width and thickness, we propose a contribution of its downregulation into dental variation. Indeed, this functional hypothesis seems to be supported by the recent association of rs198968_C as a protective allele against early childhood caries (Abbasoglu et al. 2015).

Still, we cannot discard the hypothesis of a selective advantage linked to skin traits given the documented *KLK4* expression in epidermal layers such as the *stratum granulosum*, where keratinocytes undergo dramatic morphological changes (Komatsu et al. 2003, 2005) and specially, when a *KLK4* function in epidermal pathways is supported by in vitro studies. KLK4 was not only found to activate meprin β, a protein involved in the terminal differentiation of keratinocytes (Becker-Pauly et al. 2007), as well as to trigger protease-activated receptor-2, a key mediator of melanosome transfer to keratinocytes, recurrently upregulated in darker skins (Seiberg et al. 2000a, 2000b; Sharlow et al. 2000; Seiberg 2001; Babiarz-Magee et al. 2004). Importantly, in the human lineage, skin traits linked to keratinization, epidermal differentiation, and/or pigmentation were probably under strong evolutionary pressures, due to the extensive hair loss and the subsequent need for protection against transepidermal water loss and UV skin damage (Jablonski and Chaplin 2010; Hancock et al. 2011; Gautam et al. 2015).

Consistently, many genes enrolled in epidermal pathways also show large worldwide variability and strong correlations with environmental parameters, such as solar radiation, relative humidity, and winter and summer temperatures (Hancock et al. 2011; Gautam et al. 2015). As an example, lighter skin is considered a paradigmatic case of human adaptation due to the colonization of higher latitudes and it has been correlated to polymorphism in *SLC24A5*, *SLC45A2*, *OCA2*, and *TRYP*, among others genes (McEvoy et al. 2006; Soejima and Koda 2007; Beleza et al. 2013). Nevertheless, several lines of evidence suggest a scenario of convergent evolution for the lighter skin phenotype due to a partial overlap between selected loci in European and East Asian populations (Norton et al. 2007; Berg and Coop 2014). Keratinization and epidermal differentiation traits were also recognized as targets of natural selection in humans, interconnected to climatic variables as illustrated by *EDAR* and *KRT77*, which are both engaged in thermoregulation, and also by *CDH13* and *SPINK5*, more closely related to the skin desquamation cascade in which *KLK*s are implicated (Hancock et al. 2011; Kamberov et al. 2013; Gautam et al. 2015). Particularly, KLK4 was not only described to activate KLK5, KLK7, and KLK14 present in different skin layers but also predicted to cleave several cell adhesion molecules present in the epidermis including CDH13 and desmogleins (DSG1 and DSG3) (Matsumura et al. 2005). Finally, atopic dermatitis, which has been associated with uncontrolled proteolysis, is more prevalent in northern Europe than in Asian populations, where its incidence is low (Asher et al. 2006; Naldi et al. 2009). For the reasons expressed above, it is highly attractive to speculate about a contribution of *KLK4* downregulation into East Asians skin traits. However, no epidermal phenotypes were described in dysfunctional genes (human and mouse) and protein expression in the skin still needs to be further confirmed.

Currently, we cannot rule out any of the aforementioned hypotheses, but, considering the polygenic nature of dental and skin traits, natural selection is expected to have acted simultaneously at multiple loci. Therefore, we foresee *KLK4* as another gene contributing to an adaptive phenotype possibly through a moderate effect. Recently, it has been proposed that, similar to human disease, only a fraction of selective events are monogenetic and are associated with novel mutation with large effect sizes, whereas others are polygenic adaptations, possible linked to pre-existing alleles and showing a wide range of effect sizes (Pritchard et al. 2010; Jeong and Di Rienzo 2014). Indeed, two loci (*EDAR* and *WNT10A*) were found to explain only 6% of the variance at tooth crown size in East Asians (Kimura et al. 2015) and four loci (*SLC24A5*, *SLC45A2*, *OCA2*, and *TRYP*) disclosed 35% of the variance in skin pigmentation in African-European admixed population (Beleza et al. 2013).

Finally, we further predict pleiotropic effects in male biology and other physiological functions with possible outcomes in human complex diseases. *KLK4* is a pervasive protease, expressed in a wide range of tissues, and frequently

overexpressed in prostate, ovarian and breast cancers, where it is thought to play a role in tumor progression and metastization (Kontos and Scorilas 2012). Hence, it is possible that the same haplotype conferring a selective advantage may also offer a reduced risk to several cancer types with lower incidences in East Asia (http://globocan.iarc.fr/Pages/Map.aspx, last accessed June 4, 2015) (Lose et al. 2012).

In conclusion, we confirmed the previous evidences for an adaptive evolution on the *KLK* cluster in East Asians, which is associated with a haplotype configuration acting synergistically to downregulate *KLK4*. We further propose that this haplotype was driven to high frequencies by natural selection acting on typical East Asians traits, connected either to tooth or epidermal features. Future genotype–phenotype studies would be helpful to consolidate our hypothesis.

Materials and Methods

1000 Genomes Data Set

Data from the phase I of the 1000G were retrieved from 1000 Genomes Browser website (http://browser.1000genomes.org, last accessed January 16, 2013) for the *KLK* locus (chr19:51353000–51461000, GRCh37, hg19) and for all 14 populations (1000 Genomes Project **Consortium** et al. 2010, 2012). Ancestral allele state was retrieved from Ensembl using BioMart data mining tool (http://www.ensembl.org/index.html, last accessed February 1, 2013).

Sanger Sequencing Data Set

The sequence variation of *KLK3*, *KLK2*, *KLKP1*, *KLK4*, and *KLK5* genes was surveyed in a subset of 51 samples from the DNA collection of the International HapMap Project phase I/II: 10 CEU, 11 YRI, and 30 ASN (15 CHB and 15 JPT) (supplementary table S4, Supplementary Material online).

Primers for amplification and sequencing were designed based on human genome assembly (GRCh37, hg19) for chromosome 19—nucleotide bases from 51358188 to 51447672. All samples were PCR-amplified and sequenced with BigDye Terminator v3.1 Cycle Sequencing Kit and run on an ABI 3130 automated sequencer. Details about PCR and sequencing conditions are available from authors upon request. All sequences were assembled and analyzed using Phred-Phrap-Consed package (Nickerson et al. 1997) and all putative polymorphisms were manually curated to minimize sequencing errors. Haplotypes were inferred using PHASE 2.1 (Stephens et al. 2001; Stephens and Donnelly 2003) and annotated with SNP information regarding ancestral allele state (see above).

Statistical Analysis

Summary statistics of population genetic variation for gene or sliding windows (5,000-bp window size and 1,000-bp window increment) were calculated using the online application SLIDER (http://genapps.uchicago.edu/slider/index.html). Normalized Fay and Wu's *H* (Fay and Wu 2000; Zeng et al. 2006) was calculated with ms_stats from the library molpopgen (Thornton 2003). Tajima's *D* and Fu and Li *D** statistical significance was first assessed in DnaSP v.5.10 using coalescent simulations under constant population size and no recombination (Librado and Rozas 2009). Next, for those genes with significant statistic values, we ran 100,000 coalescent simulations in ms software (Hudson 2002) under two distinct demographic models of human demography (Laval et al. 2010; Gravel et al. 2011) and assuming the recombination rate as inferred from HapMap phase II data (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/latest/old_data/rates/, last accessed November 10, 2014) (McVean et al. 2004) (supplementary table S11, Supplementary Material online).

The levels of population differentiation for ASN (CHB+JPT) versus non-ASN (YRI+CEU), ASN versus CEU, ANS versus YRI, and CEU versus YRI were calculated with the locus-by-locus F_{ST} statistic using 20,000 simulations as implemented in the Arlequin software package (Excoffier and Lischer 2010). iHS statistic was calculated and normalized for the entire chromosome 19 for the ASN population with "selscan" software (Szpiech and Hernandez 2014) after filtering SNPs with MAF less than 0.05 or with missing ancestral information state. iHS normalization was carried out using 20 equally sized allele frequency bins. Additionally, we divided chromosome 19 into nonoverlapping windows of 100 and 50 kb and calculated the proportion of SNPs with $|\mathrm{iHS}| > 2$ in each window. We applied the DIND test, which calculates the ration of ancestral to derived intrahaplotypic nucleotide diversity ($i\pi A/i\pi D$) (Barreiro et al. 2009), to the entire haplotype data of ASN population comprised within the 70-kb window defined by two major recombination hotspots. The EHH statistic (Sabeti et al. 2002) was computed with SWEEP tool (http://www.broadinstitute.org/mpg/sweep/), with cores defined as the longest nonoverlapping cores with at least three SNPs and no more than ten. Significance of REHH, given the frequency of core haplotype, was calculated in SWEEP assuming 5% frequency bins and the whole chromosome 19 data. For each core containing either rs1654556, rs198968 or rs17800874, we measured the largest genetic distance with EHH marker close to 0.05, and then considered the outmost positions of the three cores to define the largest region.

Luciferase Reporter Assay

Luciferase reporter constructs were generated for alternative allele configurations from rs198968, rs17800874, and rs1654556 SNPs. Specific haplotypes were PCR-amplified from genomic DNA of HapMap Project phaseI/II samples using the primers described in supplementary table S12, Supplementary Material online. The PCR products were first cloned into pCR2.1-TOPO vector (Life Technologies) according to manufacturer's instructions. Isolated clones were analyzed by Sanger sequencing before further processing. The constructs for rs198968 and rs17800874 were then subcloned into pGL3-Basic vector (Promega) using *Hin*dIII+*Xho*I and *Sac*I+*Xho*I (Thermo Scientific) restriction enzymes, respectively, and the constructs for rs1654556 were subcloned into pmirGLO vector (Promega) by digestion with *Sac*I and *Xho*I (Thermo Scientific). Additionally, constructs containing both rs17800874 and rs198968 were generated by subcloning the inserts for rs198968 to the 3′-end of

the rs17800874 insert using *Hin*dIII+*Xho*I restriction enzymes. Luciferase assays were carried out in three different cell lines: Human cervix adenocarcinoma (HeLa), gastric adenocarcinoma (AGS), and prostate cancer (LNCaP). HeLa cells were cultured in DMEM (Life Technologies), AGS in RPMI-1640 with Glutamax (Life Technologies), and LNCaP in RPMI-1640 medium (Life Technologies), all supplemented with 10% fetal bovine serum (Biowest) and 1% penicillin/streptomycin antibiotics (Life Technologies). HeLa and AGS cell lines were seeded at 2.0×10^5 cells per well in 24-well plate 24 h before transfection, whereas LNCaP cells were plated at 1.25×10^5 cells per well 48 h before transfection. Transient cell transfections for the empty vector (pGL3-Basic) and for the different constructs were carried out with Lipofectamine2000 (Life Technologies). A *Renilla* vector (Promega) was cotransfected and used as an internal control in transfection efficiency, and a pGL3-Control vector was used to monitor the global transfection experiment. Twenty-four hours after transfection, cells were harvested and luciferase activity was measured by Dual-Luciferase Reporter Assay System (Promega) for rs198968 and rs17800874, and by Dual-Glo Luciferase Assay System (Promega) for rs1654556 according to the manufacturer's protocol. Each experimental condition was performed in triplicate or quadruplicate and the experiments were repeated at least three times. Results are expressed as *Luciferase* activity relative to empty pGL3-basic or pmirGLO-transfected cells. Independent *t*-test was performed in GraphPad Prism 5 and the null hypothesis was rejected when P value < 0.05.

Tissue Expression Screening

To investigate the tissue specificity of *KLK3*, *KLK2*, *KLKP1*, *KLK4*, and *KLK5* expression, we analyzed 21 cDNA samples from different healthy organs derived from at least three individuals. Except for the first-strand cDNA from leukocytes (Clontech), tissue cDNA samples were synthesized by reverse transcriptase methods using as templates the RNA from the First Choice Human Total RNA Survey Panel (Ambiom). Reverse transcription was performed using the Superscript III RT-PCR system (Life Technologies) according to the manufacturer's protocol. The primers used for cDNA amplification are shown in supplementary table S13, Supplementary Material online. The amplification of a segment from *GAPDH* or *SERPINA1* was employed as internal control (Marques et al. 2013).

Supplementary Material

Supplementary figures S1–S8 and tables S1–S13 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank anonymous reviewers for their comments and helpful suggestions on the manuscript. IPATIMUP integrates the i3S Research Unit, which is partially supported by the Portuguese Foundation for Science and Technology (FCT). This work is also funded by FEDER funds through the Operational Program for Competitiveness Factors (COMPETE) and National Funds through the FCT (project PEst-C/SAU/LA0003/2013, grant PTDC/BEXGMG/0242/2012 to S.S and fellowship SFRH/BD/68940/2010 to P.I.M.) and by Programa Operacional Regional do Norte (ON.2—O Novo Norte), through FEDER funds under the Quadro de Referência Estratégico Nacional (QREN; projects NORTE-07-0124-FEDER-000024, NORTE-07-0162-FEDER-00018, and NORTE-07-0162-FEDER-000067). The authors also thank Aida Garcia and Francisco Rodriguez Diaz from University of Oviedo for kindly providing the LNCaP cell line.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.

Abbasoglu Z, Tanboga I, Kuchler EC, Deeley K, Weber M, Kaspar C, Korachi M, Vieira AR. 2015. Early childhood caries is associated with genetic variants in enamel formation and immune response genes. *Caries Res.* 49:70–77.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods.* 7:248–249.

Asher MI, Montefort S, Bjorksten B, Lai CK, Strachan DP, Weiland SK, Williams H, ISAAC Phase Three Study Group. 2006. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *Lancet* 368:733–743.

Babiarz-Magee L, Chen N, Seiberg M, Lin CB. 2004. The expression and activation of protease-activated receptor-2 correlate with skin color. *Pigment Cell Res.* 17:241–251.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, et al. 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5:e1000562.

Becker-Pauly C, Howel M, Walker T, Vlad A, Aufenvenne K, Oji V, Lottaz D, Sterchi EE, Debela M, Magdolen V, et al. 2007. The alpha and beta subunits of the metalloprotease meprin are expressed in separate layers of human epidermis, revealing different functions in keratinocyte proliferation and differentiation. *J Invest Dermatol.* 127:1115–1125.

Beleza S, Johnson NA, Candille SI, Absher DM, Coram MA, Lopes J, Campos J, Araujo, II, Anderson TM, Vilhjalmsson BJ, et al. 2013. Genetic architecture of skin and eye color in an African-European admixed population. *PLoS Genet.* 9:e1003372.

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. *PLoS Genet.* 10:e1004412.

Borgono CA, Michael IP, Diamandis EP. 2004. Human tissue kallikreins: physiologic roles and applications in cancer. *Mol Cancer Res.* 2:257–280.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15:1553–1565.

Caubet C, Jonca N, Brattsand M, Guerrin M, Bernard D, Schmidt R, Egelrud T, Simon M, Serre G. 2004. Degradation of corneodesmosome proteins by two serine proteases of the kallikrein family, SCTE/KLK5/hK5 and SCCE/KLK7/hK7. *J Invest Dermatol.* 122:1235–1244.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. *PLoS Genet.* 1:e35.

Deperthes D, Frenette G, Brillard-Bourdet M, Bourgeois L, Gauthier F, Tremblay RR, Dube JY. 1996. Potential involvement of kallikrein hK2

in the hydrolysis of the human seminal vesicle proteins after ejaculation. *J Androl.* 17:659–665.
Dixson AF. 2009. Sexual selection and the origins of human mating systems. New York: Oxford University Press Inc.
Eissa A, Diamandis EP. 2008. Human tissue kallikreins as promiscuous modulators of homeostatic skin barrier functions. *Biol Chem.* 389:669–680.
Elliott MB, Irwin DM, Diamandis EP. 2006. In silico identification and Bayesian phylogenetic analysis of multiple new mammalian kallikrein gene families. *Genomics* 88:591–599.
Emami N, Diamandis EP. 2007. Human tissue kallikreins: a road under construction. *Clin Chim Acta.* 381:78–84.
Excoffier L, Hofer T, Foll M. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb)* 103:285–298.
Excoffier L, Lischer HE. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour.* 10:564–567.
Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. 2014. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing datasets. *Mol Biol Evol.* 31(7):1850–1868.
Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
Ferreira Z, Seixas S, Andres AM, Kretzschmar WW, Mullikin JC, Cherukuri PF, Cruz P, Swanson WJ, Program NCS, Clark AG, et al. 2013. Reproduction and immunity-driven natural selection in the human WFDC locus. *Mol Biol Evol.* 30:938–950.
Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol.* 31(5):1275–1291.
Fortelny N, Cox JH, Kappelhoff R, Starr AE, Lange PF, Pavlidis P, Overall CM. 2014. Network analyses reveal pervasive functional regulation between proteases in the human protease web. *PLoS Biol.* 12:e1001869.
Fu Y. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915–925.
Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, et al. 2008. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet.* 17:835–843.
Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, Kimura R, Tokunaga K. 2008. A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Hum Genet.* 124:179–185.
Gautam P, Chaurasia A, Bhattacharya A, Grover R, Indian Genome Variation Consortium, Mukerji M, Natarajan VT. 2015. Population diversity and adaptive evolution in keratinization genes: impact of environment in shaping skin phenotypes. *Mol Biol Evol.* 32:555–573.
Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A.* 108:11983–11988.
Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327:883–886.
GTEx-Consortium. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348:648–660.
Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* 7:e1001375.
Hanihara T. 2008. Morphological variation of major human populations based on nonmetric dental traits. *Am J Phys Anthropol.* 136:169–182.
Hanihara T, Ishida H. 2005. Metric dental variation of major human populations. *Am J Phys Anthropol.* 128:287–298.
Hart PS, Hart TC, Michalec MD, Ryu OH, Simmons D, Hong S, Wright JT. 2004. Mutation in kallikrein 4 causes autosomal recessive hypomaturation amelogenesis imperfecta. *J Med Genet.* 41:545–549.
Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, 1000 Genomes Project, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331:920–924.
Hider JL, Gittelman RM, Shah T, Edwards M, Rosenbloom A, Akey JM, Parra EJ. 2013. Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evol Biol.* 13:150.
Holwerda SJ, de Laat W. 2013. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philos Trans R Soc Lond B Biol Sci.* 368:20120369.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
Jablonski NG, Chaplin G. 2010. Colloquium paper: human skin pigmentation as an adaptation to UV radiation. *Proc Natl Acad Sci U S A.* 107(Suppl 2):8962–8968.
Jeong C, Di Rienzo A. 2014. Adaptations to local environments in modern human populations. *Curr Opin Genet Dev.* 29:1–8.
Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, et al. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 152:691–702.
Kaushal A, Myers SA, Dong Y, Lai J, Tan OL, Bui LT, Hunt ML, Digby MR, Samaratunga H, Gardiner RA, et al. 2008. A novel transcript from the KLKP1 gene is androgen regulated, down-regulated during prostate cancer progression and encodes the first non-serine protease identified from the human kallikrein gene locus. *Prostate* 68:381–399.
Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* 16:980–989.
Kelley JL, Swanson WJ. 2008. Dietary change and adaptive evolution of enamelin in humans and among primates. *Genetics* 178:1595–1603.
Kimura R, Fujimoto A, Tokunaga K, Ohashi J. 2007. A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS One* 2:e286.
Kimura R, Watanabe C, Kawaguchi A, Kim YI, Park SB, Maki K, Ishida H, Yamaguchi T. 2015. Common polymorphisms in WNT10A affect tooth morphology as well as hair shape. *Hum Mol Genet.* 24:2673–2680.
Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, Haneji K, Hanihara T, Matsukusa H, Kawamura S, Maki K, et al. 2009. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am J Hum Genet.* 85:528–535.
Komatsu N, Saijoh K, Toyama T, Ohka R, Otsuki N, Hussack G, Takehara K, Diamandis EP. 2005. Multiple tissue kallikrein mRNA and protein expression in normal skin and skin diseases. *Br J Dermatol.* 153:274–281.
Komatsu N, Takata M, Otsuki N, Toyama T, Ohka R, Takehara K, Saijoh K. 2003. Expression and localization of tissue kallikrein mRNAs in human epidermis and appendages. *J Invest Dermatol.* 121:542–549.
Kontos CK, Scorilas A. 2012. Kallikrein-related peptidases (KLKs): a gene family of novel cancer biomarkers. *Clin Chem Lab Med.* 50:1877–1891.
Kraus BS. 1952. Morphologic relationships between enamel and dentin surfaces of lower first molar teeth. *J Dent Res.* 31:248–256.
Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. 2009. Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol.* 26:649–658.
Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 4:1073–1081.
Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5:e10284.
Lawrence MG, Lai J, Clements JA. 2010. Kallikreins on steroids: structure, function, and hormonal regulation of prostate-specific antigen and the extended kallikrein locus. *Endocr Rev.* 31:407–446.

Lawrence MG, Stephens CR, Need EF, Lai J, Buchanan G, Clements JA. 2012. Long terminal repeats act as androgen-responsive enhancers for the PSA-kallikrein locus. *Endocrinology* 153:3199–3210.
Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* 5:e254.
Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.
Lilja H. 1985. A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. *J Clin Invest.* 76:1899–1903.
Lose F, Srinivasan S, O'Mara T, Marquart L, Chambers S, Gardiner RA, Aitken JF, Australian Prostate Cancer B, Spurdle AB, Batra J, et al. 2012. Genetic association of the KLK4 locus with risk of prostate cancer. *PLoS One* 7:e44520.
Lovgren J, Rajakoski K, Karp M, Lundwall A, Lilja H. 1997. Activation of the zymogen form of prostate-specific antigen by human glandular kallikrein 2. *Biochem Biophys Res Commun.* 238:549–555.
Lu W, Zhou D, Glusman G, Utleg AG, White JT, Nelson PS, Vasicek TJ, Hood L, Lin B. 2006. KLK31P is a novel androgen regulated and transcribed pseudogene of kallikreins that is expressed at lower levels in prostate cancer cells than in normal prostate cells. *Prostate* 66:936–944.
Lu Y, Papagerakis P, Yamakoshi Y, Hu JC, Bartlett JD, Simmer JP. 2008. Functions of KLK4 and MMP-20 in dental enamel formation. *Biol Chem.* 389:695–700.
Lundwall A. 2013. Old genes and new genes: the evolution of the kallikrein locus. *Thromb Haemost.* 110:469–475.
Lundwall A, Brattsand M. 2008. Kallikrein-related peptidases. *Cell Mol Life Sci.* 65:2019–2038.
Marques PI, Bernardino R, Fernandes T, Program NCS, Green ED, Hurle B, Quesada V, Seixas S. 2012. Birth-and-death of KLK3 and KLK2 in primates: evolution driven by reproductive biology. *Genome Biol Evol.* 4:1331–1338.
Marques PI, Ferreira Z, Martins M, Figueiredo J, Silva DI, Castro P, Morales-Hojas R, Simoes-Correia J, Seixas S. 2013. SERPINA2 is a novel gene with a divergent function from SERPINA1. *PLoS One* 8:e66889.
Matsumura M, Bhatt AS, Andress D, Clegg N, Takayama TK, Craik CS, Nelson PS. 2005. Substrates of the prostate-specific serine protease prostase/KLK4 defined by positional-scanning peptide libraries. *Prostate* 62:1–13.
McEvoy B, Beleza S, Shriver MD. 2006. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. *Hum Mol Genet.* 15(Spec No 2):R176–R181.
McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.
Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol.* 28:659–669.
Michael IP, Pampalakis G, Mikolajczyk SD, Malm J, Sotiropoulou G, Diamandis EP. 2006. Human tissue kallikrein 5 is a member of a proteolytic cascade pathway involved in seminal clot liquefaction and potentially in prostate cancer progression. *J Biol Chem.* 281:12743–12750.
Morita W, Yano W, Nagaoka T, Abe M, Ohshima H, Nakatsukasa M. 2014. Patterns of morphological variation in enamel-dentin junction and outer enamel surface of human molars. *J Anat.* 224:669–680.
Naldi L, Parazzini F, Gallus S, Centres GS. 2009. Prevalence of atopic dermatitis in Italian schoolchildren: factors affecting its variation. *Acta Derm Venereol.* 89:122–125.
Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* 25:2745–2751.
Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD. 2007. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol Biol Evol.* 24:710–722.
Obiezu CV, Shan SJ, Soosaipillai A, Luo LY, Grass L, Sotiropoulou G, Petraki CD, Papanastasiou PA, Levesque MA, Diamandis EP. 2005. Human kallikrein 4: quantitative study in tissues and evidence for its secretion into biological fluids. *Clin Chem.* 51:1432–1442.
Olejniczak AJ, Gilbert CC, Martin LB, Smith TM, Ulhaas L, Grine FE. 2007. Morphology of the enamel-dentine junction in sections of anthropoid primate maxillary molars. *J Hum Evol.* 53:292–301.
O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res.* 18:1304–1313.
Park JH, Yamaguchi T, Watanabe C, Kawaguchi A, Haneji K, Takeda M, Kim YI, Tomoyasu Y, Watanabe M, Oota H, et al. 2012. Effects of an Asian-specific nonsynonymous EDAR variant on multiple dental traits. *J Hum Genet.* 57:508–514.
Pavlopoulou A, Pampalakis G, Michalopoulos I, Sotiropoulou G. 2010. Evolutionary history of tissue kallikreins. *PLoS One* 5:e13781.
Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20:R208–R215.
Przeworski M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* 160:1179–1189.
Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreno-Torres A, Pavlidis P, Laayouni H, Bertranpetit J, Engelken J. 2014. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.* 42:D903–D909.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.
Seiberg M. 2001. Keratinocyte-melanocyte interactions during melanosome transfer. *Pigment Cell Res.* 14:236–242.
Seiberg M, Paine C, Sharlow E, Andrade-Gordon P, Costanzo M, Eisinger M, Shapiro SS. 2000a. Inhibition of melanosome transfer results in skin lightening. *J Invest Dermatol.* 115:162–167.
Seiberg M, Paine C, Sharlow E, Andrade-Gordon P, Costanzo M, Eisinger M, Shapiro SS. 2000b. The protease-activated receptor 2 regulates pigmentation via keratinocyte-melanocyte interactions. *Exp Cell Res.* 254:25–32.
Sharlow ER, Paine CS, Babiarz L, Eisinger M, Shapiro S, Seiberg M. 2000. The protease-activated receptor-2 upregulates keratinocyte phagocytosis. *J Cell Sci. 113(Pt* 17):3093–3101.
Shaw JL, Diamandis EP. 2007. Distribution of 15 human kallikreins in tissues and biological fluids. *Clin Chem.* 53:1423–1432.
Shaw JL, Diamandis EP. 2008. Regulation of human tissue kallikrein-related peptidase expression by steroid hormones in 32 cell lines. *Biol Chem.* 389:1409–1419.
Simmer JP, Hu Y, Lertlam R, Yamakoshi Y, Hu JC. 2009. Hypomaturation enamel defects in Klk4 knockout/LacZ knockin mice. *J Biol Chem.* 284:19110–19121.
Smith P, Gomorri JM, Spitz S, Becker J. 1997. Model for the examination of evolutionary trends in tooth development. *Am J Phys Anthropol.* 102:283–294.
Soejima M, Koda Y. 2007. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. *Int J Legal Med.* 121:36–39.
Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68:978–989.

Szpiech ZA, Hernandez RD. 2014. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol.* 31:2824–2827.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Takayama TK, McMullen BA, Nelson PS, Matsumura M, Fujikawa K. 2001. Characterization of hK4 (prostase), a prostate-specific serine protease: activation of the precursor of prostate specific antigen (pro-PSA) and single-chain urokinase-type plasminogen activator and degradation of prostatic acid phosphatase. *Biochemistry* 40:15341–15348.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.

Turner CG 2nd. 1990. Major features of Sundadonty and Sinodonty, including suggestions about East Asian microevolution, population history, and late Pleistocene relationships with Australian aboriginals. *Am J Phys Anthropol.* 82:295–317.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc Natl Acad Sci U S A.* 103:135–140.

Wang SK, Hu Y, Simmer JP, Seymen F, Estrella NM, Pal S, Reid BM, Yildirim M, Bayram M, Bartlett JD, et al. 2013. Novel KLK4 and MMP20 mutations discovered by whole-exome sequencing. *J Dent Res.* 92:266–271.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7:256–276.

Yang J, Corces VG. 2012. Insulators, long-range interactions, and genome function. *Curr Opin Genet Dev.* 22:86–92.

Yoshida C, Tokumasu F, Hohmura KI, Bungert J, Hayashi N, Nagasawa T, Engel JD, Yamamoto M, Takeyasu K, Igarashi K. 1999. Long range interaction of cis-DNA elements mediated by architectural transcription factor Bach1. *Genes Cells* 4:643–655.

Yousef GM, Diamandis EP. 1999. The new kallikrein-like gene, KLK-L2. Molecular characterization, mapping, tissue expression, and hormonal regulation. *J Biol Chem.* 274:37511–37516.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174:1431–1439.

Paper III - Rare and common variants in *KLK* and *WFDC* gene families and their implications into semen hyperviscosity and other male infertility phenotypes

In preparation

Rare and common variants in *KLK* and *WFDC* gene families and their implications into semen hyperviscosity and other male infertility phenotypes

Patrícia Isabel Marques[1,2,3,4], Filipa Fonseca[1,2], Ana Sofia Carvalho[5], Diana A. Puente[3], Isabel Damião[6], Vasco Almeida[6,7], Nuno Barros[8], Alberto Barros[8,9], Filipa Carvalho[9], Rune Matthiesen[5], Victor Quesada[3], Susana Seixas[1,2]

1 - Instituto de Investigação e Inovação em Saúde, Universidade do Porto (I3S), Porto, Portugal; 2 - Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal; 3 - Department of Biochemistry and Molecular Biology-IUOPA, University of Oviedo, Oviedo, Spain; 4 - Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal; 5 – Human Genetics Department, National Institute of Health Dr Ricardo Jorge (INSA), Lisboa, Portugal; 6 – Center of Infertility and Sterility Studies (CEIE), Porto, Portugal; 7 – Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal; 8 – Center for Reproductive Genetics Alberto Barros, Porto, Portugal; 9 – Department of Genetics, Faculty of Medicine, University of Porto, Porto, Portugal.

Corresponding author: Susana Seixas
IPATIMUP, Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal
Phone: +351 225570700; Fax: +351 225570799
E-mail: sseixas@ipatimup.pt

Abstract

The human *kallikrein* (*KLK*) and whey acidic protein four-disulfide core domain (*WFDC*) gene families, located at chromosomes 19q13.3-13.4 and 20q13, respectively, encode molecules with key roles in the cascade of semen coagulation and liquefaction. Semenogelins 1 and 2 (SEMGs), the main components of the semen coagulum, establish a cross-linked matrix that entraps spermatozoa. In contrast, KLK3 and KLK2, further assisted by other KLKs, hydrolyze the coagulum in a process crucial for spermatozoa motility. Here, we examined the contribution of *KLK* and *WFDC* variation into human infertility, by performing a pilot screening of coding and non-coding regions, covering approximately 93 kb of genomic sequence, by means of a pooled sample high-throughput approach, later, followed by a genotyping survey for most promising candidates in a cohort of cases (N=238) and controls and (N=217). Among the 456 variants identified in the pooled sequencing, 296 were low-frequency for which a higher burden of deleterious alleles was detected in cases for *KLKs.* Eleven variants were confirmed to be overrepresented in cases and likely affecting KLK4 and KLK12 structure, KLK3 activity or *KLK3*, *KLK14* and *KLK15* gene expression. In SEMGs we identified 3 rare variants expected to modify the profile of cleaved peptides with potential effects in spermatozoa interactions. A common nucleotide substitution in *KLK7* (rs1654526) and a copy number variation in *SEMG1,* were on the other hand associated to a reduced risk for different infertility phenotypes. Overall, the results support the importance of *KLKs* and *SEMGs* in male reproduction and provide evidence for a contribution of their genetic variation into semen hyperviscosity and asthenozoospermia.

Introduction

Infertility is a major reproductive disorder characterized as the inability to achieve a viable pregnancy after one year of regular sexual intercourse, which affects up to 20% of couples within reproductive age in Western countries (Rowe et al. 1993; Cedenho 2007). In the disease pathogenesis, male and females are thought to have nearly equivalent contributions, as many physiological processes are required to achieve a successful fertilization. Indeed, in most couples, the two partners are often subfertile and only in a smaller percentage of cases the male factor is considered as the primary cause for a reproductive failure (Thonneau et al. 1991; Practice Committee of American Society for Reproductive 2012). So far, the recognized causes for male infertility are congenital defects leading to reproductive malformations, endocrine and immunologic dysfunction, mechanic trauma, urogenital infections causing post-testicular obstruction, impaired spermatogenesis and also major and minor genetic abnormalities, such as aneuploidies, translocations and deletions in Y and other chromosomes, as well as point mutations in genes like *CFTR*, *AR*, *DAZL*, and *SRY* (Jungwirth et al. 2012; Tahmasbpour et al. 2014). Although, infertility has been proposed as a complex disease with a strong genetic component, it continues mostly unexplained, as the large proportion of disease cases are frequently defined as idiopathic or have an unknown etiological cause (Carrell and Aston 2011; Aston 2014).

In the clinical practice, abnormal semen parameters as determined by reference threshold values of the World Health Organization (WHO), are regarded as evidence for male infertility and accordingly, infertility patients are classified into different non-mutually exclusive phenotypes. These include hyperviscosity, for a persistence of semen viscosity; oligozoospermia or azoospermia, for lower spermatozoa counts; asthenozoospermia, for reduced spermatozoa motility and teratozoospermia, for altered spermatozoa morphology (WHO 1999; WHO 2010).

Semen is a body fluid that results from an assorted mixture of the spermatozoa-rich secretions of testis and epididymis, with the products from the seminal vesicles, prostate, and bulbourethral glands. Upon ejaculation, the semen forms a gelatinous mass – the semen coagulum – that entraps the spermatozoa in a matrix of linked seminal proteins. Over a short period of time (5 to 20 minutes), the semen coagulum starts to be liquefied through the activity of several enzymes, allowing a progressive release of the spermatozoa and the regain of their motility (Lilja 1985; Deperthes et al. 1996; Robert et al. 1997).

Several proteins known to play key roles in the cascade of semen coagulation and liquefaction belong to the *kallikrein* (*KLK*) gene family, found in chromosome 19q13.3-13.4 region, and to the whey acidic protein four-disulfide core domain (*WFDC*) family, located at chromosome 20q13. Specifically, the two major structural seminal proteins enrolled in the semen coagulation and spermatozoa immobilization, semenogelin 1 (SEMG1) and semenogelin 2 (SEMG2), belong both to the *WFDC* locus (Lundwall and Brattsand 2008; Clauss et al. 2011). On the other hand, the core enzymes involved in the semen liquefaction and SEMGs hydrolysis into multiple shorter peptides are KLK3 (also known as prostate-specific antigen – PSA) and KLK2 (Lilja 1985; Lilja et al. 1987; Deperthes et al. 1996). Moreover, the *WFDC* cluster also encompasses 17 small serine protease inhibitor genes, which include SLPI and PI3, two ubiquitous molecules with antimicrobial activities at reproductive mucosal surfaces, and EPPIN, a molecule that coats the spermatozoa in a protein complex with fibronectin and SEMGs, modulates KLK3 activity and coagulum proteolysis, and also protects spermatozoa form bacterial attacks (Yenugu et al. 2004; Wang et al. 2005; Williams et al. 2006; Wang et al. 2007; Weldon et al. 2007; McCrudden et al. 2008; Zhao et al. 2008; Zhang et al. 2013). Less is known about other *WFDC* molecules, but most genes were found to be mainly expressed in male reproductive tissues and several were detected in the human seminal plasma by proteomic profiling (Clauss et al. 2002; Thimon et al. 2008; Batruch et al. 2012; Chhikara et al. 2012). Conversely, the *KLK* locus includes a total of 15 trypsin- or chymotrypsin-like serine protease genes (*KLK1-KLK15*) with pervasive activities in diverse proteolytic cascades, including semen liquefaction. As example, KLK4, KLK5, KLK14 and KLK15 were all shown to regulate the activity of KLK3, and KLK5 and KLK14 were reported to overlap with KLK3 in the hydrolysis SEMGs and fibronectin (Takayama et al. 2001a; Takayama et al. 2001b; Michael et al. 2006; Emami et al. 2008; Emami and Diamandis 2008).

In the latest years, the evidence supporting a possible contribution of KLK and WFDC families into different male infertility phenotypes has been consolidating. In the study by Emami et al., most KLKs were correlated to a downregulation of protein expression and a hyperviscosity phenotype, whereas the delay of semen liquefaction was associated to a restricted set of proteins (KLK2-3 and KLK13-14) displaying reduced protein expression (Emami et al. 2009). Furthermore, in the same study a link between KLK14 expression levels and asthezoospermia was established, and more recently KLK3 was found to be significantly upregulated in patients combining oligo and teratozoospermia phenotypes (Emami et al. 2009; Sharma et al. 2013). On the other hand, *SEMG1* has been recently shown to be upregulated in infertile patients with and without asthenozoospermia, but no correlation could be observed between the patterns of

SEMG degradation and hyperviscosity (Martinez-Heredia et al. 2008; Esfandiari et al. 2014; Legare et al. 2014; Yu et al. 2014). Moreover, in the *EPPIN* gene two single nucleotide variants (SNV) were shown to correlate with semen quality in the Han-Chinese population, one presenting a decreased risk to low sperm number and the other an increased risk to abnormal motility (Ding et al. 2010a; Ding et al. 2010b).

In this study, we sought to investigate in which extent the genetic variation within *KLK* and *WFDC* families affects the regular process of semen coagulum liquefaction, and underlies hyperviscosity and other infertility phenotypes. By performing a comprehensive survey of *KLK* and *WFDC* coding and non-coding regions using a high-throughput sequencing strategy, we demonstrated that *KLKs* have an excess of low-frequency variants among infertility cases and we validated a total of 12 candidate variants of male infertility in *KLKs* but also in *WFDCs,* with expected impact in the proteolytic processing of the semen coagulum.

Material and methods

Infertility cases and control samples

Biological samples (blood or buccal swabs and semen) from infertile men (n = 238) and controls (n=91) of Portuguese origin (2 generations at least) were collected from individuals undergoing routine spermiogram analysis at two fertility centers. All samples were classified regarding sperm count, motility and viscosity state according to the World Health Organization guidelines (WHO, 1999). Cases were classified into different infertility phenotypes and regarded as oligozoospermic, for sperm counts below 20 million/mL, as asthenozoospemic, for rapid progressive motility less than 25% and as hyperviscous, if semen drops form a thread more than 2 cm long (WHO, 1999). Patients with known causes of infertility, including chromosome anomalies and Yq microdeletions were excluded from this study. Individuals with sperm counts, motility and viscosity parameters above the mentioned thresholds were considered as normal and included in the control group. In addition, 32 individuals with at least one offspring and 94 random males with unknown spermiogram parameters, born in Portugal were included in the controls.

Genomic DNA was extracted from peripheral blood leukocytes using the Citogene Blood Kit (Citomed) or Generation Capture Column Kit (Qiagen), and from buccal swabs using BuccalAmp DNA Extraction Kit (Epicentre) according to the manufacturer's instructions. Semen samples were collected by masturbation after 3 days of sexual abstinence, centrifuged at 7000 *g* for 10 min to separate the spermatozoa from the seminal plasma and stored at -80 ºC.

Variant screening by pooled next-generation sequencing (NGS)

Initially, we carried out a pilot survey using the Somatic Mutation Identification in Pooled Samples (SMIPS) technique (Puente et al. 2011), a modified method for the analysis of pooled samples (Druley et al. 2009). In this approach (phase I), a subset of 222 samples were subdivided into three major groups, according to their phenotypes (Supplementary Fig. S1). First, as a *proxy* for a fertile group we included 57 individuals with normal sperm counts, motility and viscosity parameters and 22 men with at least one offspring. Then, the other 143 samples were further subdivided in two groups according to the presence or absence of a hyperviscosity phenotype (HV and NV, respectively), and whereas the HV group was composed by 75 cases, all showing abnormal viscosity alone or in combination with other phenotypes, the NV group consisted of the remaining 68 cases with oligozoospemia, asthenozoospermia or both, but without hyperviscosity. For

each group, we pooled equal amounts of genomic DNA per sample (>600 ng per individual, per group).

Multiple primer pairs (Supplementary Table S1) were designed based on the human genome reference sequence (GRCh37) in order to amplify 220 amplicons (119 from the *KLK* cluster and 101 from the *WFDC* cluster), covering 159 exons and 53 putative regulatory regions, according to ENCODE data for DNase hypersensivity sites, H3K4Me1 Mark and transcription factor CHIP-seq, in a total of ~ 93 kb of genomic sequence. Amplification was performed separately for each primer pair using Platinum DNA polymerase or Platinum Pfx DNA polymerase (Invitrogen/Thermo Fisher Scientific), 200 µM dNTPs (Invitrogen/Thermo Fisher Scientific), 600 nM of each primer, 1 mM $MgCl_2$, and 100 ng of pooled DNA. PCR cycling conditions were: 5 min at 94 °C, followed by 35 cycles of 30 sec at 94 °C, 30 sec at 60 °C, and 40 sec at 72 °C, and a final extension of 5 min at 72 °C. All PCR products were then purified using QIAquick PCR Purification Kit (Qiagen).

For each group, equimolecular amounts (4×10^{11} molecules) of each amplicon were pooled, and a random amplicon ligation reaction was performed as previously described (Druley et al. 2009). The concatenated products were precipitated with a 10 mM $MgCl_2$ solution, resuspended in Tris-EDTA buffer (10mM Tris; 1 mM EDTA; pH 8.0) and then fragmented in a Covaris S2 sonicator (200 Cycles, Duty Cycle 10%, intensity 5 and 360 sec) (Covaris). DNA libraries were prepared according to the Illumina protocol for paired-end sequencing, and the three libraries were run in Illumina HiSeq 2000 in a single lane (Macrogen, Inc).

All reads were initially mapped to the human reference genome (GRCh37) using the Burrows-Wheeler Alignment (BWA) tool (Li and Durbin 2009). Next, a pileup file was generated in SAMtools (Li et al. 2009) and the variant calling was performed with an in-house script and filtered based on the following criteria: i) Phred base quality score ≥ 20; ii) a minimum coverage of 10.000 reads per pool; and iii) a minimum base call count per variant corresponding to at least 0.8 expected chromosomes ((N allele-specific base calls * 2 * N individuals in pool) / N total base calls). Variant calls surpassing the filtering were normalized to allele frequencies for each pool and these were compared with publically available databases of human variation (1000 Genomes phase III) for the European population (average of the frequencies for Iberian, British, Finnish, Toscani and CEU populations - Utah residents with Northern and Western European ancestry).

Functional consequences of the identified variants were inferred using Variant Effect Predictor (VEP) tool from Ensembl, which incorporates variant location and SIFT

and Polyphen scores for changes to protein sequence (Kumar et al. 2009; Adzhubei et al. 2010; McLaren et al. 2010). The effect of synonymous substitutions in mRNA secondary structure and the outcome of non-coding variants in mRNA splicing and in the binding of transcription factors was predicted *in silico* using the online tools RNAsnp, Human Splicing Finder v.2.4.1 and MatInspector, included in the Genomatix software suite (Genomatix Software GmbH), respectively (Desmet et al. 2009; Sabarinathan et al. 2013). Structural analysis of protein sequences was done through ClustalO alignments mapped onto the three-dimensional structure of KLK4 bound to benzamidine, a competitive inhibitor (PDB ID 2BDG) (Sievers et al. 2011). The structure with the mapped residues was rendered with PyMol version 0.99rc6 (DeLano Scientific).

Genotyping validation and extension study

From the SMIPS analysis we selected several regions for variant calling validation based on the following criteria: 1) regions from genes with known implications in male reproduction, containing at least two low-frequency variants (minor allele frequency (MAF) < 0.05) with predicted functional repercussions (splice region, nonsynonymous or located in transcription factor binding regions) and displaying discrepancies between cases and controls; 2) private non-singleton low-frequency variants located in genes with possible implications in male reproduction and showing predicted functional consequences, as well as higher allele counts in cases than in controls; and 3) common SNVs (MAF ≥ 0.05) with predicted functional consequences and significant frequency differences between case and control groups. In total, 7 regions and 4 SNVs were chosen to be analyzed by Sanger sequencing or SNaPShot genotyping (phase II), respectively, in the 143 cases and 79 controls screened in phase I. All samples were PCR-amplified and whereas amplicons for *KLK3* (exon III-IV), *KLK4* (exon II), *KLK6* (exon V), *KLK12* (exons II and IV) and *KLK14* (exons I and III) were sequenced with BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific) and run on an ABI 3130 automated sequencer, the SNVs rs1654526 (*KLK7*), rs74705037 (*KLK8*), rs3212852 (*KLK15*) and rs75681320 (*EPPIN*) were genotyped using the ABI PRISM SNaPshot Multiplex Kit (Thermo Fisher Scientific) and run on an ABI 3130 automated sequencer. All Sanger sequences were assembled and analyzed using Geneious version 5.5.8 software (http://www.geneious.com, Kearse et al. 2012) and all putative polymorphisms were manually curated to minimize sequencing errors. SNaPShot peak calling was obtained and analyzed using the GeneMaper software (Thermo Fisher Scientific). Further information regarding PCR, sequencing and SNaPShot conditions are available from authors upon request. In the end, 8 SNVs were selected for an extended genotyping study (phase III) carried out in the remaining 95 infertility cases

(36 HV and 59 NV), 44 fertile controls (34 individuals with semen parameters above the threshold and 10 fertile men) and in 94 random Portuguese controls.

SEMG1 and *SEMG2* genetic screening

A pilot study covering the regions encoding the repeat units of *SEMG1* and *SEMG2* was performed by Sanger sequencing in the same subset of 222 samples used in phase I. Functional consequences of identified variants were inferred as described above. The CNV and the nonsynonymous variants predicted as damaging were further genotyped in the remaining samples by Sanger sequencing, length fragment analysis or SNaPshot genotyping. Further information regarding PCR, sequencing and SNaPShot conditions are available from authors upon request.

Statistical analyses

For the SNVs with a MAF ≥ 0.05 in the European population (1000 Genomes data phase III), statistical significant differences between cases and controls were estimated using Fisher's exact test. To control the significance levels under multiple test comparisons we performed the Bonferroni correction ($\alpha = 0.05/n$, where n is the total number of SNVs tested). Furthermore, to test for the hypothesis of a possible enrichment of low-frequency variants with predicted functional repercussions in *KLK* and *WFDC* clusters, we calculated C-alpha statistic as implemented in the AssotesteR package with 100,000 permutations (Sanchez 2013). For the statistical analysis of pooled sequenced data we performed two independent tests: one considering the nonsynonymous substitutions and splice region SNVs together, and the other one taking only into account variants located in untranslated regions (UTRs). In the analysis of *SEMG* data we carried out again two independent tests using either all nonsynonymous substitutions or only those nonsynonymous substitutions predicted as damaging. Three sets of comparisons were done throughout the study: the control group was tested against HV and NV cases separately, and a third comparison was done for the control group against all infertility cases (HV+NV). Statistical analyses were carried out using RStudio application for statistical computing (http://www.rstudio.com).

Proteomic analysis

The possible effects of KLK3 variants p.E131K and p.S210W were assessed by proteomic analysis of the seminal plasma for individuals bearing these mutations. For p.S210W the results were extracted from a seminal proteome profiling study (unpublished data). For the analysis of p.E131K, the seminal plasma was mixed with 4x Laemmli Sample Buffer (BioRad), heated at 95ºC for 5 min and separated by SDS-PAGE (12%

poly-acrylamide). The resulting gels were either stained with PageBlue, (Thermo Fisher Scientific) or transferred onto a nitrocellulose membrane (GE Healthcare) for KLK3 immunodetection. The membrane was then blocked with 5% non-fat milk and probed with the KLK3 primary antibody (SantaCruz Antibodies) at a 1:5000 dilution. Immunoblots were visualized using ECL detection kit (GE Healthcare). Protein staining band corresponding to KLK3, as assessed by immunoblotting, were manually excised from the gel, placed in 0.5 mL microcentrifuge tubes and stored at 4 °C until tryptic digestion following an already published procedures of mass spectrometry (MS) (Osorio and Reis 2013).

Results

Sequence variability of *KLK* and *WFDC* gene clusters

In order to identify potential candidate variants for male infertility among *KLK* and *WFDC* clusters, we used a pooled sample approach and a high-throughput sequencing strategy to survey a subset of 143 cases (75 HV and 68 NV) and 79 controls for 220 amplicons (119 and 101 spread over *KLK* and *WFDC* clusters, respectively) (Supplementary Fig. S1 and S2). Overall, the sequenced regions covered approximately 93 kb and included both coding and non-coding genomic segments, comprising UTRs and putative regulatory regions defined according to ENCODE data for DNase hypersensivity sites, transcription factor binding sites and histone marks.

In the data analysis, strict filtering methods were used to reduce the number of false-positives caused by sequencing errors, even so, we found an excess of novel variants in genomic segments encompassing ALUs, satellites and low complexity and simple tandem repeats. Moreover, in these regions we found large discrepancies to described SNVs frequencies for Europeans, pointing out to the existence of errors in the sequence assembling caused by the reads misalignment (Supplementary Fig. S3 and S4). These findings agreed with previous recommendations by others authors for a cautionary analysis of SNVs located in copy number variations (CNVs) and repetitive regions, due to the poor alignment of the short sequence reads (~100 bp) produced by NGS methods (Alkan et al. 2011; Treangen and Salzberg 2012).

Similar concerns were also raised in our study for *SEMGs* (Supplementary Fig. S3), more precisely, in the region encoding the repeat units of 60 amino acids (180 bp) length. Here, the high levels of sequence similarity within the variable *SEMG1* units (5 or 6 units) and between the fixed units of *SEMG2* (8 units) seem to have hampered the correct assembling of short reads to the human reference sequence. Therefore, taking into account the limitations of NGS methods in evaluating the genetic variation of repetitive elements, we decided to remove from our high-throughput analysis all SNVs identified within those regions, which included all *SEMG* amplicons.

In a further step, to assess the accuracy of our pooled sequencing approach, we compared the estimated SNVs frequencies for the control group (without novel variants) to those described for Europeans as a whole (average frequencies of populations with European ancestry from the 1000 Genomes phase III data). A strong linear correlation (r^2 = 0.826) was observed between MAFs of analyzed SNVs (Fig. 1), emphasizing the qualitative (low-frequency and common variant calling) and quantitative (allele frequency

inference) nature of the method implemented for the screening of *KLK* and *WFDC* gene variability.

Globally, in this phase of the study (phase I), a total of 456 SNVs were identified (Supplementary Table S2), in which 296 (64.9%) were low-frequency (MAF < 0.05) and 104 (22.8%) were novel variants. In addition, 98 (21.5%) SNVs were located in coding exons (58 nonsynonymous and 40 synonymous), 72 (15.8%) in UTRs and 16 (3.5%) in splice regions.

Analysis of common variants association to male infertility

To test the association to male infertility of the 160 identified common SNVs (MAF ≥ 0.05), we performed a series of comparisons between controls and different groups of cases: all infertility phenotypes (HV+NV), hyperviscosity cases (HV), or astheno- and oligozoopermic phenotypes without hyperviscosity (NV). Precisely, 29 SNVs, 23 in *KLKs* and 6 in *WFDCs*, were found to have significant associations in at least one of the comparisons (Supplementary Table S3). However, only 2 variants in the HV group maintained their statistical association after controlling for multiple test comparisons ($P < 0.0003125$). The SNV showing the strongest value ($P = 0.0002$) was a synonymous substitution in *WFDC6* (rs41304411, ENST00000372665.3:c.366C>A, p.I122I), with no obvious functional effect. The other associated SNV, rs1654526 (ENST00000595820.5:c.606+585C>T) ($P = 0.0003$), was located in *KLK7*, within a enhancer region that harbors several binding sites for transcription factors (FOS, FOSL2 and JUND) and repressors (CTCF) as indicated by chromatin segmentation and Chip-seq data from ENCODE, respectively (Supplementary Fig. S5). Interestingly, both SNVs showed higher frequencies in controls than in cases, which suggest an undirected link or a possible protective role of these SNVs to male infertility.

Burden tests of low-frequency variants

The vast majority of SNVs identified in our study were low-frequency variants (MAF < 0.05), for which standard statistic tests have low power to detect associations, especially in relative small sample sizes (few hundred individuals). To circumvent this limitation and to investigate whether there is a higher burden of low-frequency deleterious variants (nonsynonymous and splice region SNVs) in cases than in controls, we applied the C-alpha statistic under several gene combined analysis (Neale et al. 2011) (Table 1). We started by analyzing *KLK* and *WFDC* genes altogether in a single group, for which a significant enrichment of low-frequency variants in infertility cases was detected independently of the considered disease phenotype (HV, NV and HV+NV). In two other

tests, to address if *KLK* and *WFDC* clusters were differentially enriched in low-frequency SNVs, we calculated the statistic for each gene family, separately. This analysis confirmed a higher burden of low-frequency variants for all phenotypes, but solely for *KLK* genes. An equivalent approach was used to inquiry if cases had a higher burden of regulatory SNVs (5' and 3' UTR variants) than controls, nevertheless, none of the analysis yielded significant results (Table 1).

Candidate variant genotyping

To confirm variant calling and allele frequency estimates obtained in the first phase of the study (phase I), we performed a genotyping screening by Sanger sequencing for the most promising candidate genes of male infertility (phase II; Supplementary Fig. S1). Here, we prioritized the analysis of several gene regions based in the following criteria: 1) previous evidence of deregulated activity in infertility cases and/or recognized role in male reproduction; 2) gene contains at least 2 low-frequency variants with potential deleterious effects; and 3) SNV is absent in controls or displays major frequency differences between cases and controls (two or more times higher). Specifically, we selected for the Sanger sequencing study segments of *KLK3*, *KLK4*, *KLK6, KLK12* and *KLK14* genes. The function of *KLK12* has not been fully elucidated yet, but according to its mRNA expression in male genital tissues and protein identification in seminal plasma, it is likely to have an important function in male reproductive biology (Shaw and Diamandis 2007).

As a whole, the results of the Sanger sequencing screening provided a good fit to the estimated MAFs in pooled samples, as demonstrated by the strong correlations between SNVs for the three sample groups considered (HV: r^2 = 0.9725; NV: r^2 = 0.9695; controls: r^2 = 0.9509) (Fig. 2). Noticeably, in these segments the estimates of allele frequency derived from the pooled sequencing had similar levels of accuracy for both common (MAF ≥ 0.05) and low-frequency (MAF < 0.05) variants (Fig. 2).

In the category of low-frequency variants, several SNVs emerged as promising candidates for male infertility (Table 2 and Supplementary Fig. S6). In *KLK3,* we confirmed for the HV group the segregation of rs111901464 (ENST00000593997.5: c.658G>A, p.E220K), a SNV located in the intron IV of *KLK3* also predicted to have a negative impact in a shorter *KLK3* isoform; an increased incidence in HV cases of a nonsynonymous substitution (rs61729813, ENST00000326003.6: c.629C>T, p.S210W); and we identified in a single asthenozoospermic patient, a variant previously uncovered by the pooled sequencing (rs182759459, ENST00000326003.6: c.391G>A, p.E131K). While both p.E220K and p.E131K entailed a substitution of two amphipathic residues with opposite side chain charges that are frequently involved in salt-bridges and in interactions

with non-protein atoms, the p.S210W involved the change of a small polar amino acid to a large hydrophobic aromatic residue. In *KLK14*, we confirmed an increased prevalence among infertility cases of 2 deleterious SNVs, a splice donor site (rs117229324, ENST00000391802.1: c.26+1G>A) likely affecting normal splice processing (HSF and MaxEnt scores changes from 93.16 to 66.33 and 10.24 to 2.06, respectively) and a nonsynonymous variant (rs112658494, ENST00000391802.1: c.412C>T, p.R138W) replacing a positively charged amino acid by a large hydrophobic aromatic residue rarely engaged in non-protein atom binding. Furthermore, in *KLK12*, we found 2 SNVs confined to infertility cases, one identified in a single patient and causing a loss of a disulfide bond (rs140609488, ENST00000319590.8:c.587G>A, p.C196Y) and the other detected in a few cases leading to a substitution of an extremely conserved proline (rs61742847, ENST00000319590.8: c.101C>T, p.P34L) across KLK family (Supplementary Fig. S6). Finally, in *KLK4* and *KLK6*, we validated the presence of two rare deleterious variants found in isolated cases. Whereas, for KLK4, we uncovered in an asthenozoospermia case a novel p.Q42L replacement affecting a highly conserved amino acid among KLKs (Supplementary Fig. S6), in KLK6, we disclosed for a combined phenotype of asthenozoospermia and hyperviscosity a p.T234M (rs77760094, ENST00000310157.6: c.701C>T) mutation located in the C-terminal region. Noticeably, for *KLK6* we also found in two controls a novel variant p.I216N with predicted functional consequences. The p.I216N replaces a hydrophobic amino acid with an aliphatic side chain by a polar residue frequently involved in protein activity and binding sites, such as the catalytic triad of several cysteine proteases.

To investigate the putative structural consequences of the candidate variants, we mapped the affected residues onto the solved three-dimensional (3D) structure of KLK4 bound to a competitive inhibitor. Notably, the alignment of kallikrein sequences showed that several variants cluster at close or even equivalent positions in the primary structure (Fig. 3A). Moreover, the threading of the alignment onto a 3D structure revealed further clustering of positions 196 of KLK12 and 210 of KLK3 to position 215 in KLK6 (Fig. 3B). Interestingly, this cluster is located in the inhibitor-binding pocket of the structure, which suggests that these variants may directly affect the binding of the corresponding kallikreins to their substrates. In fact, the C196Y variant in KLK12 is expected to destroy a disulfide bond, termed SS6, necessary for the catalytic activity of kallikreins (Oka et al. 2002). The remaining variants map to different parts of the structure, and still show a tendency to cluster. Thus, variants E131K in KLK3 and R138W in KLK14 map to the same position in the structure (Fig. 3B). In the absence of further functional information, this suggests that variants in this position are selected for in infertile patients. Similarly,

variants Q42L in KLK4 and P34L in KLK12 cluster in consecutive positions in the structure. Both positions are well conserved among kallikreins (Fig. 3A and Supplementary Fig. S6) and appear about ten residues downstream from the activation site, suggesting that these variants might affect protein folding and the activation process. Finally, the T234M variant in KLK6 is located in a α-helix close to the C-terminus of the protein.

Beside the validation of candidate genes by Sanger sequencing 3 more dispersed variants were selected for genotyping using a SNaPShot multiplex reaction approach. These included a SNV located in the *KLK8* 5'UTR (rs74705037, ENST00000600767.5: c.-29C>T), within a region recognized by ENCODE chromatin segmentation data as an enhancer and predicted by MatInspector to contain several binding motifs for Zinc finger proteins (loss of Zinc finger and BTB domain-containing protein 7A and Zinc finger protein GLIS2 binding sites). Another SNV allocated to a *KLK15* splice region (rs3212852, ENST00000598239.5: c.481+5G>A) probably impairing normal mRNA processing (HSF and MaxEnt scores changes from 89.08 to 76.92 and 9.49 to 4.62, respectively). A latter SNV placed in *EPPIN* intron I (rs75681320, ENST00000354280.8: c.92-438C>T) in an insulator region containing several transcription binding factors (CTCF, SMC3, RAD21) as identified by ENCODE chromatin segmentation and Chip-seq data, respectively.

Extended association study of male infertility

To increase the statistic power of our study, we carried out an extended genotype analysis for the most promising candidate variants for male infertility. In this phase (phase III), we screened an additional panel of 95 cases (36 HV and 59 NV) and 138 controls (34 individuals with normal semen parameters, 10 fertile men and 94 random Portuguese males), for 7 low-frequency SNVs and a common SNV (Supplementary Fig. S1). The selected low-frequency SNVs included: 3 nonsynonymous substitutions confirmed by Sanger sequencing and predicted as deleterious (*KLK3* p.E220K - rs111901464; *KLK12* p.P34L - rs61742847 and *KLK14* p.R138W - rs112658494; the splice donor site (*KLK14*: c.26+1G>A - rs117229324); and the 3 SNVs evaluated by SNaPshot (*KLK8*: rs74705037; *EPPIN*: rs75681320 and *KLK15*: rs3212852). The single common variant selected for the extended screening was the one located on *KLK7* (rs1654526).

The case-control analysis of the entire dataset comprising a total of 238 infertility cases (111 HV and 127 NV) and 217 controls (Table 3 and Supplementary Fig. S1) corroborated a trend toward increased MAFs in infertility cases for all low-frequency variants (Table 3). However, the sample size was not enough to reach statistical significance in most circumstances. Indeed, only the *KLK12* p.P34L (rs61742847)

replacement showed a significant association for both HV+NV and HV group comparisons (*P* = 0.0388 and *P* = 0.0384, respectively), thus suggesting a possible contribution into the hyperviscosity phenotype. Nonetheless, the candidate variants identified in *KLK3*, *KLK15* and *EPPIN*, were found to present at least three times higher frequencies in the HV group than in controls. In contrary, the *KLK14* splice variant showed an equivalent frequency increment in the NV group that could be connected to an asthenozoospermia phenotype, including in the single case identified among the HV group. Finally, the common *KLK7* (rs1654526) was confirmed as significantly associated with a reduced susceptibility to semen hyperviscosity (*P* = 0.0035) and male infertility (*P* =0.0258).

SEMGs genetic screening

The survey of *SEMGs* was centered in exon II, which covers nearly all protein coding sequence, except for a short N-region included in exon I. Similarly to the approach used in the high-throughput sequencing study, we started by analyzing the same cohort of 143 cases and 79 controls by Sanger sequencing. In the first analysis of *SEMGs* variation, we identified 6 and 4 SNVs in *SEMG1* and *SEMG2*, respectively (Supplementary Table S4). Among the SNVs found in *SEMGs*, only rs147894843 (ENST00000372781.3: c.1199G>A, p.G400D) in *SEMG1,* and rs2233903 (ENST00000372769.3: c.835C>T, p.H279Y), rs2071650 (ENST00000372769.3: c.1102G>C, p.G368R) and rs139977707 (ENST00000372769.3: c.1654G>C, p.E552Q) in *SEMG2* were predicted to affect protein function. In addition, in the *SEMG1* we also covered the previously described CNV corresponding to the 5 or 6 repeat units (Jensen-Seaman and Li 2003; Lundwall et al. 2003; Miyano et al. 2003).

To investigate whether there was a higher burden of deleterious SNVs among *SEMGs* in cases than in controls, we applied again the C-alpha statistic (Neale et al. 2011) (Supplementary Table S5). However, we did not observe any variant enrichment in all tests performed. In this initial phase, a single silent substitution located in *SEMG1* and restricted to controls (p.T293T rs17850164) was found to displayed a significant association (*P* = 0.0423, in the HV+NV comparison). Nevertheless, 2 nonsynonymous substitutions were absent in controls (SEMG1 p.G400D and SEMG2 p.E552Q); 2 linked variants in *SEMG2* (p.H279Yand p.G368R) were slightly increased in cases, and the 5 repeat allele of SEMG1 showed a higher frequency in controls than in cases. Interestingly, the significantly associated p.T293T mutation had no predicted effect on mRNA secondary structure but it appeared as linked to the 5-repeat allele. Thus, in the extended study, we decided to genotype for *SEMG1* the p.G400D replacement and the CNV; and for *SEMG2* the p.H279Y and p.E552Q substitutions.

The analysis of the full case-controls datasets revealed a significant association of *SEMG1* p.G400D (rs147894843) variant in HV+NV and NV groups (*P* = 0.0388 and *P* = 0.04920, respectively) (Table 4). This variant creates a potential cleavage site (P1-P1': DE search in MEROPS) for cathepsin D and metalloprotease 2, two proteases found at high and low abundances in semen, respectively and also for caspase-3, a cysteine protease previously associated to male infertility and asthenozoospermia (Almeida et al. 2005; Rawlings et al. 2014). Consistently, in our study 4 out of 5 cases carrying the p.G400D had an asthenozoospermia phenotype, for which a significant association was also obtained (*P* = 0.0442, asthenozoospermia *vs*. controls). The *SEMG2* variants did not reach significance in our group comparisons, nevertheless, we found a slighter increment on the p.H279Y (rs2233903) replacement in the NV group and a 7 times augmented frequency of p.E552Q (rs139977707) in the HV. Oddly, this SNV is likely to generate a novel cleavage site for *KLK3* (P1-P1': QS) in the SEMG2 sequence (Malm et al. 2000).

The CNV maintained the tendency toward a lower frequency of the 5-repeat allele in infertile patients (*P* = 0.0667 for HV+NV cases *vs.* controls), as previously reported in Asians (Miyano et al. 2003). Notably, in our sample, we could detect a significant lowering of the 5-repeat allele in oligozoospermia patients (*P* = 0.02928) but not in asthenozoospermia as it has been previously hypothesized (Miyano et al. 2003).

At last, in the extended survey of *SEMGs* we also discovered in a single individual exhibiting a combined phenotype of hyperviscosity and asthenozoospermia a novel SEMG1 variant (hg19 chr20: g.43837061T>C, p.Y315H), expected to abolish one of the KLK3 cleavage sites (Rawlings et al. 2014).

Proteomic validation

To assess a possible deleterious effect of the candidate variants that could result in the degradation of the mutant protein, we carried out a proteomic screening of KLK3 p.E131K and p.S210W substitutions in the seminal plasma of individuals bearing these mutations. In both cases, it was possible to identify the mutant variant in comparable levels to those for the wild type (Fig. 4 for p.S210W; for p.E131K data is not shown), indicating that the mutant allele is secreted and not degraded by the cellular machinery. However, other deleterious functional effect of these variants in KLK3 function cannot be excluded: the p.S210W substitution by being placed in the binding pocket can still affect the interactions to substrates, and in the p.E131K the loss of glutamate residue could potentially affect the binding to ions like zinc, an important modulator of KLK activity in the seminal plasma.

Discussion

We performed a comprehensive study of male infertility focused on the genetic variation of *KLK* and *WFDC* clusters, by applying a next-generation sequencing approach to the analysis of pooled DNA samples, allowing in a cost-effective manner to identify multiple susceptibility markers in cases with and without hyperviscosity. A follow-up genotype screening of selected locus regions confirmed the accuracy of the method, which permitted the variant calling and frequency inference of both common and rare alleles.

Globally, an enrichment of potential candidate variants (high and low-frequency) was detected in *KLKs* in comparison to *WFDCs*, supporting a greater impact of the first cluster in human reproduction and fertility. Furthermore, in agreement with *KLKs* higher levels of sequence variation in infertility cases, other authors had reported a consistent downregulation of protein expression for most tested KLKs (KLK1-3, KLK5-10 and KLK13-14) in the semen of individuals with abnormal viscosity and liquefaction parameters (Emami et al. 2009).

Although, a greater emphasis was given in our study, in particular in the genotyping surveys, to nonsynonymous and splice variants, these may only explain a small fraction of KLK quantitative differences in semen through mechanisms of abnormal protein and mRNA degradation. Indeed, among validated variants only the p.P34L substitution in KLK12, which is located in a highly conserved residue across the entire family, as well as, in other far related serine proteases, and the mutated donor splice region of *KLK15,* could hypothetically be linked to hyperviscosity by those straightforward effects. However, neither KLK12 nor KLK15 were previously evaluated in the semen, and whereas the later protease is known to activate pro-KLK3 *in-vitro*, a role of *KLK12* in male reproduction was not yet been elucidated (Takayama et al. 2001a). Worth to note, in this study two common variants in *KLK12* region that did not surpass multiple testing were also associated to the hyperviscosity phenotype.

On the other hand, SNVs positioned in regulatory regions may also cause significant differences in protein expression, if these disrupt a binding motif for an important transcription factor. Even though, this category of genetic variants of male infertility has been less explored in the genotyping phases of the study because of the less confident nature of bioinformatic predictions, several candidates are likely to have their link to hyperviscosity explained by such phenomena. This is likely to be the case of the low-frequency variant in the 5’ UTR of *KLK8* (rs74705037), slightly augmented in hyperviscosity cases and the significantly associated variant, located in an intron of KLK7

(rs1654526), both placed in enhancer regions previously shown by ENCODE to interact with multiple regulatory elements. Interestingly, KLK7 besides being found in extremely reduced concentrations in hyperviscosity cases (Emami et al. 2009), its gene harbors three other common SNVs less strongly associated to this phenotype (rs1991820, rs1991819, and rs1991818).

Another genotyped SNV possibly correlated to the hyperviscosity phenotype through gene downregulation is a low-frequency variant in KLK3 (rs111901464) positioned in a region identified by ENCODE chromatin segmentation as weaker enhancer, but that could also affect a shorter protein isoform lacking the catalytic triad (p.E220K). Up till now, a reduced expression of KLK3 has only been described in cases of delayed viscosity and displaying reduced spermatozoa number with abnormal morphology (Emami et al. 2009; Sharma et al. 2013).

Two other nonsynonymous variants, predicted as deleterious, were found to be slightly increased in hyperviscosity (near two times higher), the KLK3 p.S210W (rs61729813) and the KLK14 p.R138W (rs112658494). However, these SNVs if truly linked to infertility are only expected to cause qualitative changes in protein interactions with other molecules. Specifically, the p.S210W variant of KLK3 was confirmed to not differentially affect the protein content in the semen and since it is located in a balcony region of the catalytic pocket (Ser210-Gly225) exposed to bulk solvents probably affects protease activity (Debela et al. 2006). Here, a substitution of a small polar residue (serine) by a large aromatic amino acid (tryptophan), where other KLK often present a glutamine (Q), is likely to restrain the substrate acceptance in the KLK3 catalytic pocket.

Despite the limited number of SNVs identified in the *WFDC* cluster, two possible susceptibility variants were discovered to be more prevalent among the hyperviscosity cases, a SNV located in a regulatory region, specifically in an insulator of *EPPIN*, a smaller protease inhibitor reported to target KLK3, also implicated in antimicrobial activities (McCrudden et al. 2008) and a p.E552Q substitution in SEMG2. This later variant is likely to introduce a novel cleave site for KLK3 in the C-terminal region of the protein, but so far no specialized function has been attributed yet to such region (Robert et al. 1997; Robert and Gagnon 1999).

Still, other susceptibility variants evaluated in the genotyping screening were overrepresented among infertility cases with normal viscosity. These included a nonsynonymous variant located in a critical structural region, p.Q42L in KLK4, and the disrupted donor splice of KLK14, all possible contributing to a lowering of the KLK content. Conversely, the remaining identified variants are mostly likely affecting molecular

interactions or protein activity, namely the KLK3 p.E131K, the KLK12 p.C196Y and the SEMG1 substitutions p.Y315H and p.G400D. Notably, both SEMG1 variants are expected to alter the protein proteolytic processing, but only the later substitution is in close proximity to a key sequence recognized as a thyrotropin-releasing hormone (TRH) like peptide (375-397 residues). Despite some controversy about the SEMG1 origin of the TRH-like peptides found in the human semen, these were demonstrated to increase the capacitation of spermatozoa (Khan and Smyth 1993; Huber et al. 1998; Robert and Gagnon 1999). More recently, nearly the same sequence of SEMG1 (376-388 residues) was found to bind to the CD52 glycosylphosphatidylinositol anchored antigen presented by spermatozoa, which also takes part in semen coagulation and its released during liquefaction (Flori et al. 2008).

Most of these susceptibility SNVs found in normal viscosity cases were otherwise correlated in several instances to asthenozoospermia, a finding that may be still consistent with abnormal patterns of semen liquefaction. SEMGs and specially SEMG1 are described to bound spermatozoa and to modulate their motility in a dose- and time dependent fashion (Robert and Gagnon 1996; Yoshida et al. 2008; Mitra et al. 2010). So far, the motility inhibitory peptides were correlated with N-terminal peptides released during semen liquefaction (α-inhibin-92 and α-inhibin-31) and containing cysteine 239 residue, but the functional relevance of other SEMG1 regions should not be discarded (Silva et al. 2013). There is a growing body of evidence for the interaction of SEMG1 with other biomolecules, in most cases, with functional outcomes in spermatozoa motility. SEMG1 has been reported as a target for S-nitroso-glutathione, and for prolactin inducible protein, to bind zinc and while in EPPIN-complexes to interact with spermatozoa calcium channels (Lefievre et al. 2007; Yoshida et al. 2008; O'Rand and Widgren 2012; Tomar et al. 2013). Conversely, in asthenozoospermia patients SEMG1 was shown to remain bound to spermatozoa, to be increased in their semen, and its mRNA to be highly expressed by spermatozoa (Zhao et al. 2007; Martinez-Heredia et al. 2008; Terai et al. 2010; Yu et al. 2014).

In overview, all identified variants are expected to have a negative impact in fertility through modifications of semen liquefaction process, resulting from a deregulation of protein/peptide activity and concentration levels. Still, the consequences in the proteolytic cascade of each SNV might be difficult to disentangle, given that KLKs are known to have pervasive links with each other, often activating other members and overlapping in substrate affinity (Lawrence et al. 2010). Furthermore, low-frequency variants were always found in single heterozygosity cases, which indicate that in those individuals they are only

playing a part into the genetic background of male infertility. Moreover, candidate SNVs are only present in small fraction of cases and therefore, far from fully explaining KLK and SEMG association to hyperviscosity and asthenozoospermia phenotypes. This finding is concordant with a view of male infertility as a complex disease resulting from a complex network of genetic and non-genetic factors with a wide range of susceptibility effects.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248-249.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12:363-376.

Almeida C, Cardoso MF, Sousa M, Viana P, Goncalves A, Silva J, Barros A. 2005. Quantitative study of caspase-3 activity in semen and after swim-up preparation in relation to sperm quality. Hum Reprod 20:1307-1313.

Aston KI. 2014. Genetic susceptibility to male infertility: news from genome-wide association studies. Andrology 2:315-321.

Batruch I, Smith CR, Mullen BJ, Grober E, Lo KC, Diamandis EP, Jarvi KA. 2012. Analysis of seminal plasma from patients with non-obstructive azoospermia and identification of candidate biomarkers of male infertility. J Proteome Res 11:1503-1511.

Carrell DT, Aston KI. 2011. The search for SNPs, CNVs, and epigenetic variants associated with the complex disease of male infertility. Syst Biol Reprod Med 57:17-26.

Cedenho AP. 2007. Evaluation of the subfertile male. Pp. 115-140. Male Infertility.

Chhikara N, Saraswat M, Tomar AK, Dey S, Singh S, Yadav S. 2012. Human epididymis protein-4 (HE-4): a novel cross-class protease inhibitor. PLoS One 7:e47672.

Clauss A, Lilja H, Lundwall A. 2002. A locus on human chromosome 20 contains several genes expressing protease inhibitor domains with homology to whey acidic protein. Biochem J 368:233-242.

Clauss A, Persson M, Lilja H, Lundwall A. 2011. Three genes expressing Kunitz domains in the epididymis are related to genes of WFDC-type protease inhibitors and semen coagulum proteins in spite of lacking similarity between their protein products. BMC Biochem 12:55.

Debela M, Magdolen V, Schechter N, Valachova M, Lottspeich F, Craik CS, Choe Y, Bode W, Goettig P. 2006. Specificity profiling of seven human tissue kallikreins reveals individual subsite preferences. J Biol Chem 281:25678-25688.

Deperthes D, Frenette G, Brillard-Bourdet M, Bourgeois L, Gauthier F, Tremblay RR, Dube JY. 1996. Potential involvement of kallikrein hK2 in the hydrolysis of the human seminal vesicle proteins after ejaculation. J Androl 17:659-665.

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67.

Ding X, Zhang J, Bian Z, Xia Y, Lu C, Gu A, Li Y, Song L, Wang S, Wang X. 2010a. Variants in the Eppin gene show association with semen quality in Han-Chinese population. Reprod Biomed Online 20:125-131.

Ding X, Zhang J, Fei J, Bian Z, Li Y, Xia Y, Lu C, Song L, Wang S, Wang X. 2010b. Variants of the EPPIN gene affect the risk of idiopathic male infertility in the Han-Chinese population. Hum Reprod 25:1657-1665.

Druley TE, Vallania FL, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS et al. 2009. Quantification of rare allelic variants from pooled genomic DNA. Nat Methods 6:263-265.

Emami N, Deperthes D, Malm J, Diamandis EP. 2008. Major role of human KLK14 in seminal clot liquefaction. J Biol Chem 283:19561-19569.

Emami N, Diamandis EP. 2008. Human kallikrein-related peptidase 14 (KLK14) is a new activator component of the KLK proteolytic cascade. Possible function in seminal plasma and skin. J Biol Chem 283:3031-3041.

Emami N, Scorilas A, Soosaipillai A, Earle T, Mullen B, Diamandis EP. 2009. Association between kallikrein-related peptidases (KLKs) and macroscopic indicators of semen analysis: their relation to sperm motility. Biol Chem 390:921-929.

Esfandiari N, de Lamirande E, Gukturk A, San Gabriel MC, Nazemian Z, Burjaq H, Casper RF, Zini A. 2014. Seminal hyperviscosity is not associated with semenogelin degradation or sperm deoxyribonucleic acid damage: a prospective study of infertile couples. Fertil Steril 101:1599-1603.

Flori F, Ermini L, La Sala GB, Nicoli A, Capone A, Focarelli R, Rosati F, Giovampaola CD. 2008. The GPI-anchored CD52 antigen of the sperm surface interacts with semenogelin and participates in clot formation and liquefaction of human semen. Mol Reprod Dev 75:326-335.

Huber AE, Fraser H, del Rio-Garcia J, Kreil G, Smyth DG. 1998. Molecular cloning in the marmoset shows that semenogelin is not the precursor of the TRH-like peptide pGlu-Glu-Pro amide. Biochim Biophys Acta 1387:143-152.

Jensen-Seaman MI, Li WH. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. J Mol Evol 57:261-270.

Jungwirth A, Giwercman A, Tournaye H, Diemer T, Kopa Z, Dohle G, Krausz C, European Association of Urology Working Group on Male I. 2012. European Association of Urology guidelines on Male Infertility: the 2012 update. Eur Urol 62:324-332.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649.

Khan Z, Smyth DG. 1993. Isolation and identification of N-terminally extended forms of 5-oxoprolylglutamylprolinamide (Glp-Glu-Pro-NH2), a thyrotropin-releasing-hormone (TRH)-like peptide present in human semen. Eur J Biochem 212:35-40.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073-1081.

Lawrence MG, Lai J, Clements JA. 2010. Kallikreins on Steroids: Structure, Function, and Hormonal Regulation of Prostate-Specific Antigen and the Extended Kallikrein Locus. Endocr Rev 31:407-446.

Lefievre L, Chen Y, Conner SJ, Scott JL, Publicover SJ, Ford WC, Barratt CL. 2007. Human spermatozoa contain multiple targets for protein S-nitrosylation: an alternative mechanism of the modulation of sperm function by nitric oxide? Proteomics 7:3066-3084.

Legare C, Droit A, Fournier F, Bourassa S, Force A, Cloutier F, Tremblay R, Sullivan R. 2014. Investigation of male infertility using quantitative comparative proteomics. J Proteome Res 13:5403-5414.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Lilja H. 1985. A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. J Clin Invest 76:1899-1903.

Lilja H, Oldbring J, Rannevik G, Laurell CB. 1987. Seminal vesicle-secreted proteins and their reactions during gelation and liquefaction of human semen. J Clin Invest 80:281-285.

Lundwall A, Brattsand M. 2008. Kallikrein-related peptidases. Cell Mol Life Sci 65:2019-2038.

Lundwall A, Giwercman A, Ruhayel Y, Giwercman Y, Lilja H, Hallden C, Malm J. 2003. A frequent allele codes for a truncated variant of semenogelin I, the major protein component of human semen coagulum. Mol Hum Reprod 9:345-350.

Malm J, Hellman J, Hogg P, Lilja H. 2000. Enzymatic action of prostate-specific antigen (PSA or hK3): substrate specificity and regulation by Zn(2+), a tight-binding inhibitor. Prostate 45:132-139.

Martinez-Heredia J, de Mateo S, Vidal-Taboada JM, Ballesca JL, Oliva R. 2008. Identification of proteomic differences in asthenozoospermic sperm samples. Hum Reprod 23:783-791.

McCrudden MT, Dafforn TR, Houston DF, Turkington PT, Timson DJ. 2008. Functional domains of the human epididymal protease inhibitor, eppin. FEBS J 275:1742-1750.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069-2070.

Michael IP, Pampalakis G, Mikolajczyk SD, Malm J, Sotiropoulou G, Diamandis EP. 2006. Human tissue kallikrein 5 is a member of a proteolytic cascade pathway involved in seminal clot liquefaction and potentially in prostate cancer progression. J Biol Chem 281:12743-12750.

Mitra A, Richardson RT, O'Rand MG. 2010. Analysis of recombinant human semenogelin as an inhibitor of human sperm motility. Biol Reprod 82:489-496.

Miyano S, Yoshida K, Yoshiike M, Miyamoto C, Furuichi Y, Iwamoto T. 2003. A large deletion of the repeat site in semenogelin I is not involved in male infertility. Int J Mol Med 11:435-440.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. PLoS Genet 7:e1001322.

O'Rand MG, Widgren EE. 2012. Loss of calcium in human spermatozoa via EPPIN, the semenogelin receptor. Biol Reprod 86:55.

Oka T, Hakoshima T, Itakura M, Yamamori S, Takahashi M, Hashimoto Y, Shiosaka S, Kato K. 2002. Role of loop structures of neuropsin in the activity of serine protease and regulated secretion. J Biol Chem 277:14724-14730.

Osorio H, Reis CA. 2013. Mass spectrometry methods for studying glycosylation in cancer. Methods Mol Biol 1007:301-316.

Practice Committee of American Society for Reproductive M. 2012. Diagnostic evaluation of the infertile male: a committee opinion. Fertil Steril 98:294-301.

Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, Escaramis G, Jares P, Bea S, Gonzalez-Diaz M et al. 2011. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 475:101-105.

Rawlings ND, Waller M, Barrett AJ, Bateman A. 2014. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 42:D503-509.

Robert M, Gagnon C. 1996. Purification and characterization of the active precursor of a human sperm motility inhibitor secreted by the seminal vesicles: identity with semenogelin. Biol Reprod 55:813-821.

Robert M, Gagnon C. 1999. Semenogelin I: a coagulum forming, multifunctional seminal vesicle protein. Cell Mol Life Sci 55:944-960.

Robert M, Gibbs BF, Jacobson E, Gagnon C. 1997. Characterization of prostate-specific antigen proteolytic activity on its major physiological substrate, the sperm motility inhibitor precursor/semenogelin I. Biochemistry 36:3811-3819.

Rowe PJ, Comhaire FH, Hargreave TB, Mellows HJ. 1993. WHO Manual for the Standardized Investigation and Diagnosis of the Infertile Couple. Cambridge University Press.

Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J. 2013. RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. Hum Mutat 34:546-556.

Sanchez G. 2013. Package 'AssotesteR'.

Sharma R, Agarwal A, Mohanty G, Jesudasan R, Gopalan B, Willard B, Yadav SP, Sabanegh E. 2013. Functional proteomic analysis of seminal plasma proteins in men with various semen parameters. Reprod Biol Endocrinol 11:38.

Shaw JL, Diamandis EP. 2007. Distribution of 15 human kallikreins in tissues and biological fluids. Clin Chem 53:1423-1432.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.

Silva EJ, Hamil KG, O'Rand MG. 2013. Interacting proteins on human spermatozoa: adaptive evolution of the binding of semenogelin I to EPPIN. PLoS One 8:e82014.

Tahmasbpour E, Balasubramanian D, Agarwal A. 2014. A multi-faceted approach to understanding male infertility: gene mutations, molecular defects and assisted reproductive techniques (ART). J Assist Reprod Genet 31:1115-1137.

Takayama TK, Carter CA, Deng T. 2001a. Activation of prostate-specific antigen precursor (pro-PSA) by prostin, a novel human prostatic serine protease identified by degenerate PCR. Biochemistry 40:1679-1687.

Takayama TK, McMullen BA, Nelson PS, Matsumura M, Fujikawa K. 2001b. Characterization of hK4 (prostase), a prostate-specific serine protease: activation of the precursor of prostate specific antigen (pro-PSA) and single-chain urokinase-type plasminogen activator and degradation of prostatic acid phosphatase. Biochemistry 40:15341-15348.

Terai K, Yoshida K, Yoshiike M, Fujime M, Iwamoto T. 2010. Association of seminal plasma motility inhibitors/semenogelins with sperm in asthenozoospermia-infertile men. Urol Int 85:209-215.

Thimon V, Calvo E, Koukoui O, Legare C, Sullivan R. 2008. Effects of vasectomy on gene expression profiling along the human epididymis. Biol Reprod 79:262-273.

Thonneau P, Marchand S, Tallec A, Ferial ML, Ducot B, Lansac J, Lopes P, Tabaste JM, Spira A. 1991. Incidence and main causes of infertility in a resident population (1,850,000) of three French regions (1988-1989). Hum Reprod 6:811-816.

Tomar AK, Sooch BS, Raj I, Singh S, Yadav S. 2013. Interaction analysis identifies semenogelin I fragments as new binding partners of PIP in human seminal plasma. Int J Biol Macromol 52:296-299.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13:36-46.

Wang Z, Widgren EE, Richardson RT, O'Rand MG. 2007. Characterization of an eppin protein complex from human semen and spermatozoa. Biol Reprod 77:476-484.

Wang Z, Widgren EE, Sivashanmugam P, O'Rand MG, Richardson RT. 2005. Association of eppin with semenogelin on human spermatozoa. Biol Reprod 72:1064-1070.

Weldon S, McGarry N, Taggart CC, McElvaney NG. 2007. The role of secretory leucoprotease inhibitor in the resolution of inflammatory responses. Biochem Soc Trans 35:273-276.

WHO. 1999. WHO Laboratory Manual for the Examination of Human Semen and Sperm-Cervical Mucus Interaction - 4th edition.

WHO. 2010. WHO laboratory manual for the Examination and processing of human semen - 5th edition.

Williams SE, Brown TI, Roghanian A, Sallenave JM. 2006. SLPI and elafin: one glove, many fingers. Clin Sci (Lond) 110:21-35.

Yenugu S, Richardson RT, Sivashanmugam P, Wang Z, O'Rand M G, French FS, Hall SH. 2004. Antimicrobial activity of human EPPIN, an androgen-regulated, sperm-bound protein with a whey acidic protein motif. Biol Reprod 71:1484-1490.

Yoshida K, Kawano N, Yoshiike M, Yoshida M, Iwamoto T, Morisawa M. 2008. Physiological roles of semenogelin I and zinc in sperm motility and semen coagulation on ejaculation in humans. Mol Hum Reprod 14:151-156.

Yu Q, Zhou Q, Wei Q, Li J, Feng C, Mao X. 2014. SEMG1 may be the candidate gene for idiopathic asthenozoospermia. Andrologia 46:158-166.

Zhang X, Fang J, Xu B, Zhang S, Su S, Song Z, Deng Y, Wang H, Zhao D, Niu X et al. 2013. Correlation of epididymal protease inhibitor and fibronectin in human semen. PLoS One 8:e82600.

Zhao C, Huo R, Wang FQ, Lin M, Zhou ZM, Sha JH. 2007. Identification of several proteins involved in regulation of sperm motility by proteomic analysis. Fertil Steril 87:436-438.

Zhao H, Lee WH, Shen JH, Li H, Zhang Y. 2008. Identification of novel semenogelin I-derived antimicrobial peptide from liquefied human seminal plasma. Peptides 29:505-511.

Table 1 – Burden tests for *KLK* and *WFDC* low-frequency variants.

		Cases (HV+NV) ***vs.*** **Controls**	**HV Cases** ***vs*** **Controls**	**NV Cases** ***vs.*** **Controls**
Nonsynonymous and splice region SNVs	*KLKs* and *WFDCs*	C-alpha = 50.9067 (*P*-value = <u>0.0327</u>)	C-alpha = 30.7813 (*P*-value = <u>0.0034</u>)	C-alpha = 22.9367 (*P*-value = <u>0.0080</u>)
	WFDCs	C-alpha = 9.1536 (*P*-value = 0.2872)	C-alpha = -1.1391 (*P*-value = 0.8730)	C-alpha = 6.5306 (*P*-value = 0.1536)
	KLKs	C-alpha = 41.7530 (*P*-value = <u>0.0106</u>)	C-alpha = 31.9205 (*P*-value = <u>0.0001</u>)	C-alpha = 16.4061 (*P*-value = <u>0.0066</u>)
UTR SNVs	*KLKs* and *WFDCs*	C-alpha = -11.0989 (*P*-value = 0.5599)	C-alpha = -2.0320 (*P*-value = 0.8988)	C-alpha = -9.2201 (*P*-value = 0.5252)
	WFDCs	C-alpha = 7.2740 (*P*-value = 0.0696)	C-alpha = 5.0348 (*P*-value = 0.0654)	C-alpha = 3.4013 (*P*-value = 0.1822)
	KLKs	C-alpha = -18.3729 (*P*-value = 0.1808)	C-alpha = -7.0668 (*P*-value = 0.6191)	C-alpha = -12.6214 (*P*-value = 0.2766)

Significant *P*-values ($P < 0.05$) are underlined

Table 2 – Low-frequency variants surveyed in phase II.

Gene	SNP ID	Genomic position (hg19)	Consequence	MAF (number of chromosome)			
				Control	Cases (HV+NV)	HV cases	NV cases
KLK3	rs182759459	51361469	p.E131K	0.000 (0)	0.003 (1)	0.000 (0)	0.007 (1)
KLK3	rs61729813	51361850	p.S210W	0.013 (2)	0.014 (4)	0.020 (3)	0.007 (1)
KLK3	rs111901464	51361879	intronic*	0.000 (0)	0.010 (3)	0.020 (3)	0.000 (0)
KLK4	rs138071534	51412573	p.E53E	0.000 (0)	0.003 (1)	0.000 (0)	0.007 (1)
KLK4	N/A	51412607	p.Q42L	0.000 (0)	0.003 (1)	0.000 (0)	0.007 (1)
KLK4	N/A	51412717	intronic	0.000 (0)	0.003 (1)	0.007 (1)	0.000 (0)
KLK4	N/A	51412740	intronic	0.000 (0)	0.003 (1)	0.007 (1)	0.000 (0)
KLK6	rs77760094	51462454	p.T234M	0.000 (0)	0.003 (1)	0.007 (1)	0.000 (0)
KLK6	N/A	51462508	p.I215N	0.013 (2)	0.000 (0)	0.000 (0)	0.000 (0)
KLK8	rs74705037	51504808	5'UTR	0.006 (1)	0.035 (10)	0.033 (5)	0.037 (5)
KLK12	rs140609488	51534048	p.C196Y	0.000 (0)	0.003 (1)	0.000 (0)	0.007 (1)
KLK12	rs61742847	51537332	p.P34L	0.000 (0)	0.010 (3)	0.013 (2)	0.007 (1)
KLK14	rs112658494	51582808	p.R138W	0.006 (1)	0.028 (8)	0.040 (6)	0.015 (2)
KLK14	rs117229324	51585978	splice donor	0.006 (1)	0.017 (5)	0.007 (1)	0.029 (4)
KLK15	rs3212852	51330129	splice region	0.013 (2)	0.017 (5)	0.027 (4)	0.007 (1)
EPPIN	rs75681320	44174847	intronic	0.013 (2)	0.024 (7)	0.040 (6)	0.007 (1)

* p.E220K predicted as damaging in a shorter isoform. N/A – Not applicable, novel variants without ID.

Table 3 – Combined case-control association analysis from phase III.

Gene	SNP ID	Consequence	Control	Cases (HV+NV)		HV cases		NV cases	
			MAF	MAF	*P*-value	MAF	*P*-value	MAF	*P*-value
KLK3	rs111901464	Intronic	0.002	0.008	0.2170	0.014	0.1150	0.004	0.6001
KLK7	rs1654526	Intronic	0.267	0.210	0.0258	0.171	0.0035	0.246	0.3019
KLK8	rs74705037	5'UTR	0.018	0.025	0.3207	0.027	0.3236	0.024	0.4120
KLK12	rs61742847	p.P34L	0.000	0.011	0.0388	0.014	0.0384	0.004	0.3673
KLK14	rs112658494	p.R138W	0.014	0.019	0.3686	0.027	0.1860	0.012	0.5653
KLK14	rs117229324	splice donor	0.005	0.011	0.2653	0.005	0.7343	0.016	0.1362
KLK15	rs3212852	splice region	0.007	0.019	0.0965	0.023	0.0919	0.016	0.2284
EPPIN	rs75681320	Intronic	0.009	0.019	0.1713	0.027	0.0803	0.012	0.5077

Significant nominal *P*-values ($P < 0.05$) are underlined.

Table 4 – Combined case-control association analysis for SNVs in *SEMG1* and *SEMG2*.

Gene	SNP ID	Consequence	Control	Cases (HV+NV)		HV cases		NV cases	
			MAF	MAF	*P*-value	MAF	*P*-value	MAF	*P*-value
SMEG1	rs147894843	p.G400D	0.000	0.011	0.0388	0.009	0.1142	0.012	0.0492
SMEG1	5 repeat units	del Ia (aa320-379)	0.047	0.032	0.0667	0.027	0.1355	0.028	0.1313
SMEG2	rs2233903	p.H279Y	0.014	0.021	0.2856	0.014	0.6383	0.028	0.1581
SMEG2	rs139977707	p.E552Q	0.002	0.006	0.3475	0.014	0.1150	0.000	0.6327

Significant nominal *P*-values ($P < 0.05$) are underlined.

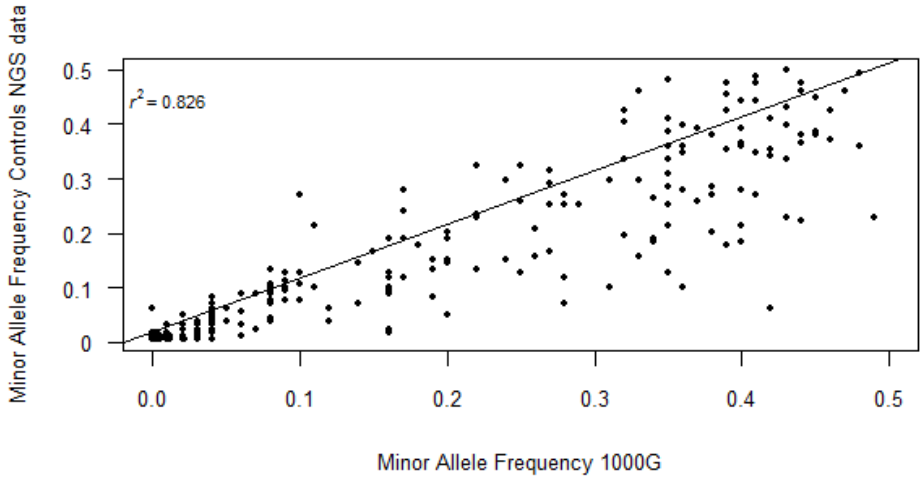


Figure 1 – Minor allele frequencies (MAFs) from 1000 Genomes data *vs.* controls from pooled sequencing. Allele frequency estimates for 277 SNVs based on pooled sequencing from the control group were compared with the described European average frequencies from 1000 Genomes project phase III samples. r^2 - correlation coefficient (r^2 = 0.826).

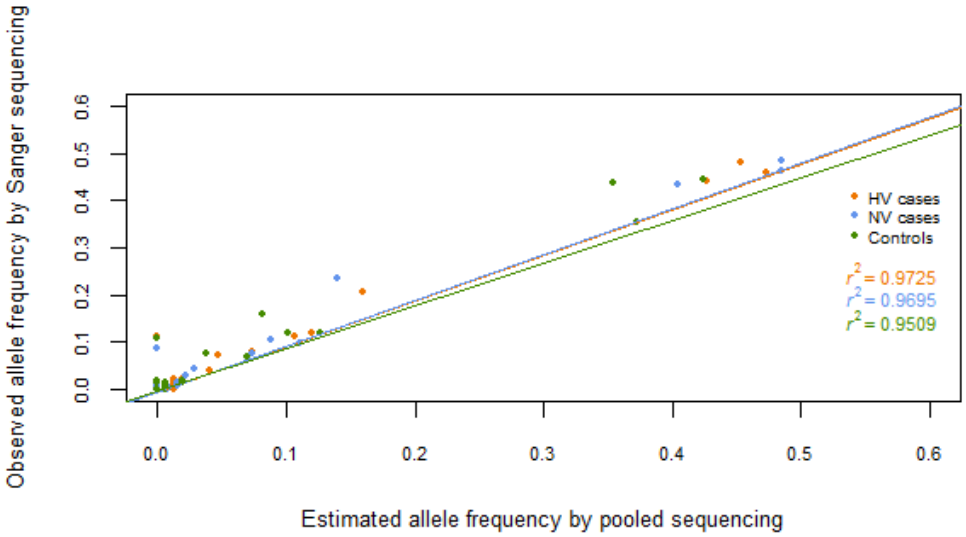


Figure 2 – Minor allele frequencies (MAFs) from pooled sequencing *vs.* Sanger sequencing. Estimated MAFs based on pooled sequencing is plotted against the actual frequencies as determined by individual Sanger sequencing for the surveyed regions. r^2 - correlation coefficients (HV: r^2 = 0.9725; NV: r^2 = 0.9695; controls: r^2 = 0.9509). The data from HV cases, NV cases and controls are represented in orange, blue and green, respectively.

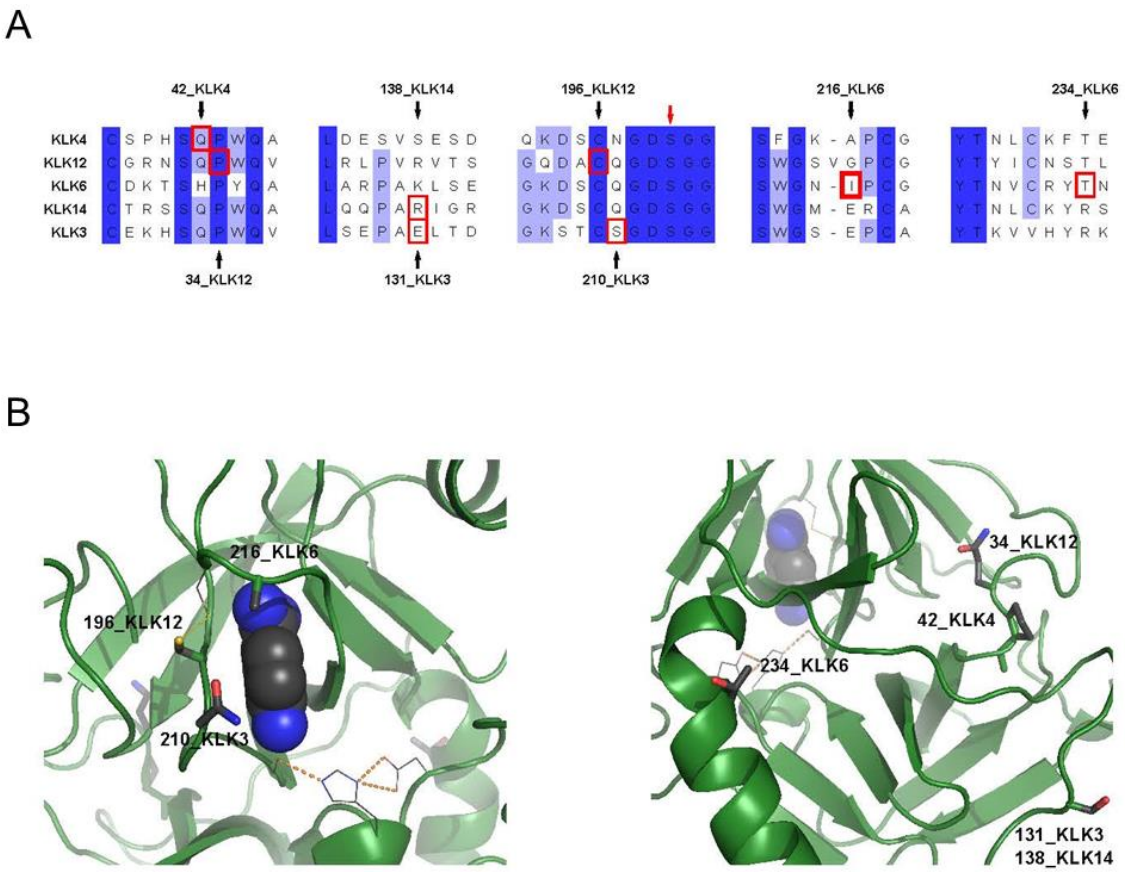


Figure 3 – Structural characterization of the KLK low-frequency variants. (A) Alignment of the amino acid sequences of the variant kallikreins. Variant sites are framed in red. Complete conservation is shown in dark blue background, whereas partial conservation is shown on a light blue background. The catalytic serine is highlighted with a red arrow. **(B)** Mapping of variant sites on a kallikrein structure. The overall structure is depicted as a green ribbon. Variant sites are shown as sticks. The catalytic triad and the second SS6 cysteine are shown as lines.

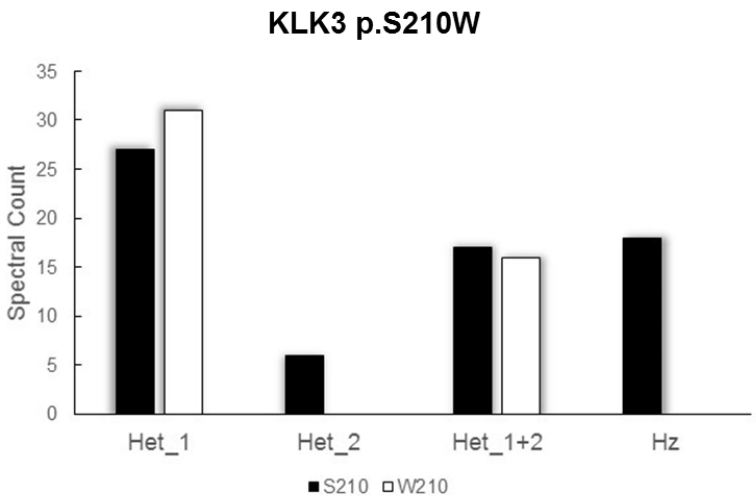


Figure 4 – Relative abundance of KLK3 p.S210W variant in seminal plasma. Spectral counts for p.S210W residue in two heterozygous (Het_1, Het_2) and of 33 homozygous (Hz) individuals. Total spectral counts are shown for Het_1 and Het_2 individuals, and the mean of spectral counts are displayed for Het_1+2 and Hz.

Chapter 4

Final Discussion

Natural selection is an important process by which populations adapt and evolve. Among the plethora of targets of natural selection, several classes of reproductive-related genes, including those encoding for seminal plasma proteins, were shown to be under positive selection and variable selective pressures within and between lineages (Jensen-Seaman and Li 2003; Dorus et al. 2004; Clark and Swanson 2005; Hurle et al. 2007; Carnahan and Jensen-Seaman 2008; Ramm et al. 2008; Ferreira et al. 2013a; Ferreira et al. 2013b). Considering the importance of *KLKs* in many essential biological processes, especially in reproduction, this work aimed to elucidate the extent of natural selection on these genes, at inter and intra-specific level, and to uncover genetic variants with potential implications in health and disease.

1. Evolutionary history of *KLKs*

Gene duplication has long been recognized as a major source of raw genetic material for evolutionary innovation (Ohno 1970). According to the "birth-and-death" evolution hypothesis, new genes are generated by duplication and, while some are maintained in the genome by acquiring new and altered functions or differentiated spatial and temporal expression, others are deleted or become non-functional through disruptive mutations (Nei and Rooney 2005; Eirin-Lopez et al. 2012). This model is thought to be more common among genes underlying large variable physiological traits across species, such as those involved in sensory perception, immunity and reproduction. Indeed, the evolution of several reproductive-related genes was shown to be greatly marked by these processes (Tian et al. 2009). Two remarkable examples are found in genes encoding the zona pellucida (ZP) proteins, the extracellular matrix surrounding oocytes and in a transglutaminase (*TGM4*), which is recognized to be involved in the formation of semen coagulum and copulatory plugs in several primates and rodents species (Clark and Swanson 2005; Goudet et al. 2008; Tian et al. 2009). In both instances, several losses of *ZP* and *TGM4* genes were found throughout the vertebrate evolution, whereas for *ZP* genes, it was hypothesized that these may be partially associated with the transition from an external to internal fertilization (from anamniotes to amniotes), for *TGM4* these seemed to be related with lack of semen coagulation (opossum - *Monodelphis domestica*, cow-*Bos Taurus*, gibbon and gorilla).

In the first part of this work, the evolution of *KLK2* and *KLK3*, the most recent duplicates of the *KLK* cluster with known key roles in semen liquefaction, was addressed by analyzing a panel of 22 primate species (Paper I). Overall, the disclosed evolutionary

history of these two genes is consistent with the “birth-and-death” evolution model: (1) *KLK3* was originated by a duplication of *KLK2*, an event that occurred after the Catarrhini split approximately 42 mya; (2) *KLK3* functionally diverged from *KLK2* mainly through the acquisition of a D207S substitution, which resulted in a change of the substrate affinity (trypsin- to chymotrypsin-like activity) and an extended spectrum of cleavage sites in SEMGs; and (3) *KLK2* gene was inactivated in several species by different genomic mechanisms including gene deletion, pseudogenization and unequal crossing-over. The latter phenomenon was connected to independent events of *KLK3-KLK2* gene fusion giving rise to chimeric *KLKs* (*cKLK*) in gorilla and northern white-cheeked gibbon (*Nomascus leucogenys*) (Paper I), which could be reconciled with previous reports of a partial *KLK2* loss in gorilla and gibbon species (Clark and Swanson 2005). Given that, at protein level, these rearrangements account only for amino acid substitutions not predicted to affect KLK3 structure or function, it was advanced that these *cKLKs* were likely to be functionally active (Paper I). The analysis of the seminal plasma of these taxa could easily corroborate such hypothesis, however, the characterization of the ejaculate proteome of non-human primates is still restricted to a limited number of species, in which, unfortunately, gorilla and gibbon were not included (Claw 2013). Similarly, evidence of *cKLK* transcripts could indicate that, at least, these are transcribed, but the screening of *KLK3* mRNA in prostate tissues from non-human primates also did not comprise the species in which these rearrangements occur (Mubiru et al. 2014).

In the seminal proteins, adaptive evolution has been hypothesized to result from several major driving forces including sperm-egg recognition, male attempts to counteract the female immune response, host-pathogen interactions and post-copulatory selection (Clark and Swanson 2005; Ramm et al. 2008; Dorus et al. 2010; Findlay and Swanson 2010). In the latter case, sperm competition has been recognized to affect several physiological and morphological traits in males, and its intensity to be highly correlated with mating behavior in primates (Harcourt et al. 1981; Dixson and Anderson 2002; Anderson et al. 2007; Dixson 2009). For example, multimale species exhibit larger testis relative to body size, higher sperm counts, bigger sperm midpiece volume, greater mitochondria load and prominent semen coagulation. Importantly, SEMGs have been previously identified as genes under adaptive evolution, in which their evolutionary rates and number of repeat units were found to correlate to different mating systems and to the rate of semen coagulation across several primate species (Jensen-Seaman and Li 2003; Dorus et al. 2004; Hurle et al. 2007). Notably, the number of functional KLKs was significantly associated with the content of SEMG repeats and primate mating systems (Paper I). On one hand, active KLK2 and KLK3 were correlated to higher SEMG repeat

numbers and a multimale mating behavior, whereas, on the other hand, the lack of one or both proteins was linked to lower repeat numbers and a unimale mating system, suggesting that the evolution of these two *KLK* genes might have been driven by reproductive biology, in a sperm competition manner and in a close relation to SEMGs. Consistently, a recent study centered in the evolution of 1170 seminal proteins detected an overrepresentation of pseudogenes within unimale lineages in comparison to multimale taxa (Claw 2013). In agreement with these findings, the loss of KLK2 in primates was more frequently observed in unimale than in multimale species (7 unimale and 2 multimale) (Paper I).

Recently, two studies aiming to estimate the evolutionary rates of seminal and non-seminal proteins in hominoids, by the usage of comprehensive gene datasets, showed that the concept of reproductive genes as fast evolving genomic regions is indeed an oversimplification, since this broad category of genes tends to be under stronger selective pressures than those from non-reproductive tissues (Good et al. 2013; Carnahan-Craig and Jensen-Seaman 2014). Furthermore, the same authors also reported that in spite of some reproductive proteins being probable targets of adaptive evolution driven by sexual selection forces, in general, seminal proteins do not present faster evolutionary rates in species with strong sperm competition. Moreover, these authors also proposed that the evolutionary rates are more correlated to function and tissue specificity, rather than to the mating system (Good et al. 2013; Carnahan-Craig and Jensen-Seaman 2014). Still, the detection of post-copulatory selection may not be so straightforward due to the complexity of protein networks and proteolytic cascades, as illustrated by the results presented here for *KLK2* and *KLK3* (Paper I)(Fortelny et al. 2014).

A possible role of *KLKs* in human adaptation is also plausible if one considers the coevolution of *KLK2* and *KLK3* with SEMGs in primates (Paper I) and a previously reported signature of natural selection for *SEMG1* in Asian populations, which was correlated to an altered proteolytic profile and antimicrobial activities in semen (Ferreira et al. 2013b). Indeed, hallmarks of natural selection in the *KLK2-KLK5* locus for East Asians were detected by three GWS of positive selection, based on distinct statistics and databases of human genetic variation, but those have remained unexplored (Voight et al. 2006; Pickrell et al. 2009; Pybus et al. 2014). To better understand the evolutionary forces acting at the *KLK3-KLK5* segment in East Asians, a comprehensive analysis of *KLK* genetic diversity in the region was carried out using mostly 1000G phase I data (Paper II).

In general, the observed patterns of genetic diversity for *KLK3-KLK5* region, as summarized by Tajima's *D*, Fu and Li *D** and Fay and Wu's *H* statistics, suggested an excess of both low-frequency variants and high-frequency derived alleles in East Asians.

According to recent evidence, these results could simply be attributed to the explosive growth of human populations in the last 10,000 years, which is thought to have caused an increase of rare variants (Gravel et al. 2011; Keinan and Clark 2012). Nevertheless, the statistical significance obtained in the most updated models of Asian demography, already assuming a large population expansion, seems to favors a selective hypothesis (Laval et al. 2010; Gravel et al. 2011; Keinan and Clark 2012).

Even though the *KLK* region could be associated to several statistical arguments for a non-neutral evolution, the selective footprints did not fit a standard selective sweep, as they were centered in a relatively short segment spanning from *KLK2* to *KLK5* (~70 kb) and encompassing multiple recombination hotspots (Paper II). Such findings were not surprising given the current understanding of the field, which proposes classic selective sweeps as rare events in human evolution and the current patterns of human genetic diversity as an end product of milder selective events (soft sweeps or selection on standing variation) (Pritchard and Di Rienzo 2010; Pritchard et al. 2010; Hernandez et al. 2011; Crisci and Jensen 2012; Granka et al. 2012).

Moreover, recent investigations estimated that only ~0.5% of nonsynonymous substitutions have been targeted by natural selection in the last 250,000 years and many human adaptations are instead related to expression quantitative trait loci (eQTL) (Hernandez et al. 2011; Vernot et al. 2012; Grossman et al. 2013; Jha et al. 2015). In agreement with such assumption, none of the nonsynonymous SNPs at the *KLK2-KLK5* locus fitted a positive selection hypothesis and, likewise, the most promising candidate variants (rs1654556_G, rs198968_T and rs17800874_A) lay in putative regulatory regions in *KLK4* and in the *KLK4-KLK5* intergenic sequence (Paper II). These variants defined a common haplotype (GTA) in East Asians, which according to functional *in vitro* assays and *in vivo* data may downregulate *KLK4* expression (Paper II)(GTEx-Consortium 2015). Remarkably, rs198968_T and rs17800874_A were shown to operate synergistically and to have a much higher effect than rs1654556_G in *KLK4* expression, pointing out the former variants as the main *KLK4* eQTLs where rs1654556_G might only play a minor contribution to gene downregulation.

Although the GTA haplotype in other populations was only present at very low frequencies, or nearly absent, the three variants can be found outside Asia at intermediate haplotype configurations (GCA, GTG, GCG, ACA and ATG), which may suggests that the selected haplotype originated by recombination and then was swept to high frequencies only in East Asians (Figure 8). The recently released 1000G phase III data, which contains novel populations with South Asian, East Asian and African ancestries, as well as, more individual samples for the populations already screened in phase I, shows the

same discrepancies in allele and haplotype frequencies, further indicating the East Asian nature of the selective signature (Genomes Project et al. 2015).

It is important to stress out that while rs1654556_G represents the ancestral allele state, variants rs198968_T and rs17800874_A are both derived alleles forms, and all of them might have been neutral before the emergence of the GTA haplotype, as indicated by their low frequencies in non-East Asian populations. Therefore, this may fit better a scenario of selection on standing variation (soft sweep), in which variants were already segregating in the population before the selective pressures had favored an increment in variant frequency (Hermisson and Pennings 2005; Przeworski et al. 2005; Pritchard and Di Rienzo 2010; Peter et al. 2012; Messer and Petrov 2013).

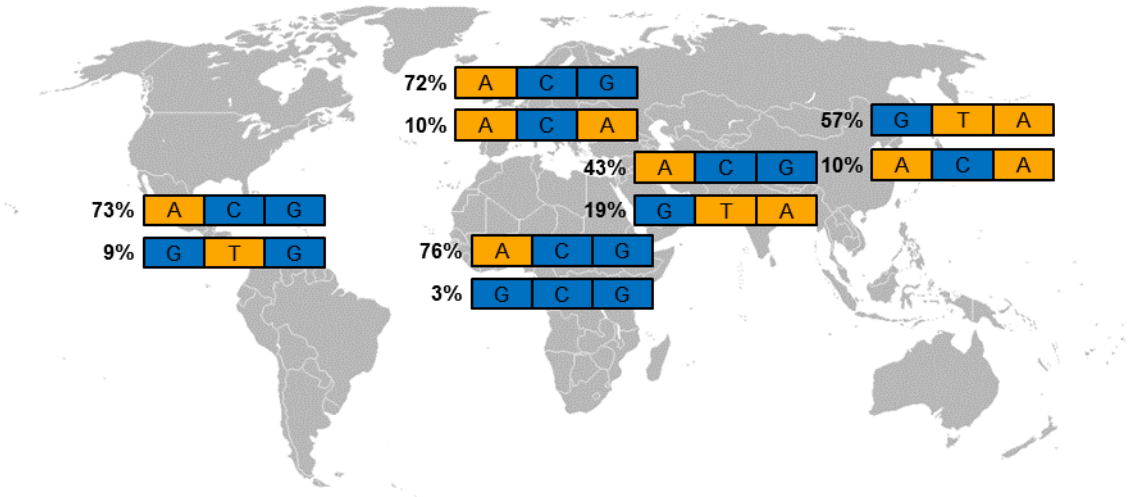


Figure 8 – Worldwide estimated haplotype frequencies defined by rs1654556, rs198968 and rs17800874 according to 1000G phase III data for African, European, South Asia, East Asia and American populations. For each continental region the most common haplotypes are shown. In Africa the ancestral haplotype is also displayed. Ancestral and derived alleles are represented in blue and orange, respectively.

As long as humans spread out of Africa they experienced a whole new set of climate variables and changes in food availability, which are likely to have dictated different subsistence strategies (Hancock et al. 2010a; Hancock et al. 2010b). In this regard, genetic variation already standing in populations may have played a prominent role in human evolutionary history since it would contribute to a faster adaptive response

to alterations in selective pressures. Indeed, such fast response patterns are observed in the case of artificial selection in maize (Innan and Kim 2004; Durand et al. 2010).

A selective hypothesis associated with reproductive functions was considered at fist given that *KLK4* has been also proposed to have a role in the proteolytic cascade of semen liquefaction and considering the reported importance of *KLKs* in the evolution of semen coagulation in primates (Paper I) (Takayama et al. 2001b). However, male physiological traits indicate that sperm competition is unlikely to have played a part in human evolution and these are rather consistent with a history of monogamy (unimale mating systems) (Dixson 2009). For example, in comparison with other mammals, humans have small testis relative to body size, longer spermatogenesis times, lower sperm cell counts and tinier spermatozoa midpieces. Furthermore, it was hypothesized that the reduced testis size observed in some modern Asian populations is not due to sperm competition, neither to changes in their mating system, but instead, are a by-product of the selective forces acting on gonadal function in women to prevent multiple ovulation (Dixson 2009). Therefore, taking into account the traits known to have been targeted by natural selection in human populations and the possible implications of *KLK4* in these processes, two selective hypothesis related to tooth and epidermal features were advanced (Paper II).

On one hand, the expression of *KLK4* in some epidermal layers together with its ability to activate *in vitro* several molecules involved in important pathways of skin physiology, such as keratinization, melanosome transfer and skin desquamation, turned out to be an extremely attractive hypothesis (Seiberg et al. 2000; Komatsu et al. 2003; Babiarz-Magee et al. 2004; Komatsu et al. 2005; Matsumura et al. 2005; Becker-Pauly et al. 2007; Ramsay et al. 2008). Although, protein expression in skin is still a controversial issue and the *KLK4* inactivation in *amelogenesis imperfecta* disease and mouse models does not seem to have major outcomes in epidermal phenotypes (Hart et al. 2004; Simmer et al. 2009; Wang et al. 2013). Nonetheless, it is important to note that skin as well as many human morphological traits, is thought to be polygenic, in which KLK4 may only contribute through a moderate effect and, hence, subtle changes in phenotypes might have remained undetected.

On the other hand, *KLK4* play an essential role in tooth maturation and accordingly, its disrupted activity result in enamel defects in humans and mice (Hart et al. 2004; Simmer et al. 2009; Wang et al. 2013). Additionally, *KLK4* arose by a duplication of *KLK5* near the divergence of Boreoeutheria (e.g. primates, rodents and artiodactyls) and Afrotheria (e.g. elephant, hyrax and tenrec) lineages, where the latter taxonomic group presents a characteristic delayed in dental eruption (Kawasaki et al. 2014). Moreover,

KLK4 has been pseudogenized in toothless minke and bowhead whales (*Balaenoptera acutorostrata* and *Balaena mysticetus*, respectively), whereas in the toothed cetaceans (orca - *Orcinus orca* and bottlenose dolphin - *Tursiops truncates*) the gene remained intact (Kawasaki et al. 2014; Keane et al. 2015). In humans, and in East Asians in particular, it is possible to observe common dental variations, like upper central incisor shoveling, enamel extensions of the first maxillary molar and other dental traits, often designated as sinodonty, which can be partially associated with the genetic variation in *EDAR*, another gene showing footprints of positive selection in East Asians (Turner 1990; Hanihara and Ishida 2005; Sabeti et al. 2007; Hanihara 2008; Kimura et al. 2009; Kamberov et al. 2013).

For the reasons mentioned above, and bearing in mind that tooth morphologies are also likely a polygenic trait, *KLK4* might represent yet another gene contributing to an adaptive tooth phenotype through a moderate effect. Still, the pervasive nature of *KLK4*, its wide pattern of tissue expression and the observed outcomes of selected variants in three different cellular systems, seem to suggest the existence of pleiotropic effects in other physiological functions, which may result in increased resistance or susceptibility to human disease.

2. Implication of *KLKs* genetic variation in human health and disease

Most, if not all, biological processes in humans comprise complex proteolytic networks, in which spatial and temporal alterations can lead to different disease states (Lopez-Otin and Bond 2008; Quesada et al. 2009). In this scope, the *KLK* gene family is no exception and despite the numerous studies performed already to understand the impact of *KLKs* in both normal and pathological conditions, significant pieces of information are still missing. For instance, from a clinical perspective, most of the literature has been focused on cancer risk, where *KLK3* is by far the best studied member of the family, due to its early recognition as a biomarker for prostate cancer. Nevertheless, the recent development of diverse *Klk* mouse models (knockout, knock-in and transgenic) have shed light into the associations of each *KLK* with distinctive human disorders, such as *amelogenesis imperfecta*, atopic dermatitis, myocardial ischaemia, multiple sclerosis and schizophrenia (Ny and Egelrud 2004; Pons et al. 2008; Simmer et al. 2009; Smith et al. 2011; Tamura et al. 2012; Murakami et al. 2013; Furio et al. 2014). Nonetheless, the potential implications of *KLK* variation outside of malignancy remain largely unknown.

Hence, in-depth comprehensive surveys of the *KLK* locus within varied diseases, including male infertility, are fundamental.

In a third work, the contribution of *KLKs* variability was explored in the framework of male infertility and three non-mutually exclusive phenotypes, hyperviscosity, asthenozoospermia and oligozoospermia, simultaneously with the study of their substrates and inhibitors, the *SEMGs, EPPIN and EPPIN-like* genes, all belonging to the *WFDC* family located at chromosome 20q13 (Paper III).

Male infertility is considered a multifactorial disease with a strong genetic component, in which common variants may account with small increments into disease susceptibility (Carrell and Aston 2011). In this particular disorder, it is rather straightforward that strong deleterious alleles will be rapidly wipe out from extant human populations because of their negative impact in the reproductive fitness. Conversely, as mentioned earlier, as human populations have expanded many novel variants were originated and, due to their recent origins, they will still be segregating at low-frequencies in populations in spite of a possible influence in human fertility (codominant inheritance). Altogether, male infertility, or more precisely male subfertility, like other human complex disorders, can be considered as a consequence of a wide range of genetic variants with variable size effects into the different disease phenotypes (Aston and Carrell 2009; Aston et al. 2010; Carrell and Aston 2011; Lettre 2014). Specifically, in the case-control study of male infertility centered on the *KLK* and *WFDC* loci, the majority of identified candidate variants with potential serious effects in protein structure, activity and interactions, splicing processing and expression regulation were low-frequency variants, among which *KLKs* were overrepresented, suggesting a greater burden of these genes in male reproduction in comparison to *WFDCs* (Paper III).

Among the recognized infertility phenotypes, delayed liquefaction, hyperviscosity and asthenozoospermia are the ones more probably correlated to a deregulated activity of KLKs and WFDCs in the semen, given that a small modification of the fine-tuned protein interactions in the semen might potentially interfere in its quality. Consistently, most KLKs have been previously described to be downregulated in individuals exhibiting the hyperviscosity phenotype (KLK1-2, KLK5-8, KLK10, KLK13-14) and a few KLKs were also found to display a significantly reduced protein expression in cases with delayed liquefaction (KLK2-3 and KLK13-14) (Emami et al. 2009). On the other hand, associations between abnormal sperm motility (asthenozoospermia) and KLK14 downregulation, SEMG1 augmented expression and the higher prevalence of *EPPIN* low-frequency variants were also reported (Martinez-Heredia et al. 2008; Emami et al. 2009; Ding et al. 2010a; Ding et al. 2010b). In overview, the *KLK* and *WFDC* genetic variants identified in

Paper III comprised several SNVs possibly influencing the protein contents in the semen by mechanisms of abnormal protein and mRNA degradation (nonsynonymous substitutions and splice variants) or by gene misregulation (variants located in regulatory regions), which could in some instances relate to the reduced KLK levels already reported. Conversely, several other variants (nonsynonymous substitutions only) were found to likely affect important protein interactions in semen because of the insertion or removal of proteolysis cleavage sites, or even the substrate allowance in the protease catalytic pocket, all of them possibly connected to hyperviscosity and asthenozoospermia phenotypes. Furthermore, the distribution of these susceptibility markers across *KLKs* seems to suggest *KLK3*, *KLK12* and *KLK14* as the genes with greater impact in male fertility.

Although a greater focus was given to nonsynonymous replacements and splice variants in the genotyping phases, mainly due to the reduced reliability of bioinformatics predictions for nucleotide substitutions located in transcription factor binding sites, the variants placed at coding regions may only explain a minor part of the genetic predisposition to male infertility (Paper III). Indeed, only a small fraction of the 456 identified variants were nonsynonymous substitutions (12.7%; 9.6% low-frequency), which reinforces a possible contribution of gene regulation in the disease pathogenesis. This finding concurs with previous assumptions made for targets of natural selection, in which eQTLs are also thought to play a significant role in complex diseases by altering the gene expression timings and intensities (Epstein 2009; Nicolae et al. 2010; Bryois et al. 2014). In addition, it has been shown that common SNVs associated with complex traits are more likely to be eQTLs than nonsynonymous variants, in which the effect size is inversely correlated to allele frequency, in brief, the greater the allele frequency, the lower the contribution to the trait (Nicolae et al. 2010; Battle et al. 2014).

Interestingly, for a common variant located *KLK7* and a low-frequency CNV of *SEMG1*, both showing significant associations with male infertility, the ancestral alleles were the ones conferring an increased susceptibility to the disease. From an evolutionary perspective, this evidence fits well the “thrifty genotype” theory, which proposes alleles that were once beneficial in past human history, nowadays are less adapted to the environmental conditions of western societies (Neel 1962; Neel et al. 1998). Consistently, other complex disorders linked to this theory, like obesity and diabetes, are known to affect human fertility, as well as lifestyles habits such as smoking and alcohol consumption and tendency for parents having the first child at increased ages (Vine et al. 1994; Sharpe and Franks 2002).

Finally, the semen is a complex body fluid known to be enriched in proteases and these have been established to have pervasive links with each other, their inhibitors and substrates (Pilch and Mann 2006; Batruch et al. 2011; Laflamme and Wolfner 2013; Fortelny et al. 2014). Therefore, it is attractive to speculate that some of these enzymes may have critical roles in the semen liquefaction profiling with possible functional repercussions in the spermatozoa abilities to fertilize the egg. Globally, these *KLK* and *WFDC* variants are far from explaining the genetic basis of male infertility, or even hyperviscosity or asthenozoospermia phenotypes, and only represent a minor proportion of its susceptibility and deleterious variation. In the near future, with the continuous decline of whole exome and genome sequencing costs, together with the advances in proteome technologies, new insights concerning the pathogenesis of male infertility will probably emerge with benefits in patient personalized medicine.

Chapter 5

Concluding Remarks

This work provides further support for an implication of proteolysis genes, specifically *KLKs*, as targets of adaptive evolution in primates through diverse biological functions including male reproduction and for a possible role of *KLK* sequence variation in both beneficial traits and disease phenotypes.

1. It sustains the hypothesis that the evolution of *KLK2* and *KLK3* might have been driven by reproductive biology in a sperm competition manner. This study contributed not only to consolidate the origin of *KLK3* in Catarrhini through an event of duplication and functional divergence from *KLK2* towards a chymotrypsin-like activity and, consequently, an extended enzymatic spectrum over SEMG1 and SEMG2 cleavage, but also to unveil a complex dynamics of *KLK2* and *KLK3* evolution mediated by different genomic mechanisms of gene loss in a close correlation to SEMG gene structure, primate mating system and semen coagulation rates.

2. It clarifies a signal of natural selection shaping the genetic diversity of the *KLK* cluster in Asian populations. A complex signature of positive selection was disclosed in East Asians favoring a haplotype defined by three variants (rs1654456_G, rs198968_T and rs17800874_A) acting synergistically to downregulate *KLK4*. A reasonable hypothesis is that this haplotype was driven to high frequencies in East Asians by offering a selective advantage into phenotypic traits characteristic of these populations, such as tooth shape and structure, and epidermal features. Still, due to the pervasive nature of *KLK4*, the possibility of pleiotropic effects in other physiological functions, including in reproductive biology, also exists and may result in an increased disease resistance or susceptibility. In a wider perspective, this study endorses the concept that soft sweeps and polygenic adaptation may actually be more common in human evolution than classic selective sweeps.

3. It highlights several low-frequency variants at *KLK* and *SEMG* genes that may contribute to different male infertility phenotypes. An excess of low-frequency variants was observed for the *KLK* cluster but not for the *WFDC* locus and this burden was independent of the infertility phenotype considered. Among the most promising variants, two were found to overlap onto the protein structure in KLK3 (p.E131K) and in KLK14 (rs112658494, p.R138W). Furthermore, one variant in *KLK12* (rs61742847, p.P34L) and another in *SEMG1* (rs147894843, p.G400D) were significantly associated with semen hyperviscosity and asthenozoospermia, respectively. In addition, a common

intronic variant in *KLK7* (rs1654526) and the copy number variation embracing one of the repeat units in *SEMG1* were connected to a protective role against hyperviscosity and oligozoospermia, respectively. Nevertheless, the identified candidate variants are expected to only play a part in the genetic background of male infertility, given that they were always found in heterozygosity and that KLKs often overlap in substrate affinity.

Chapter 6

References

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19:711-722.

Anderson MJ, Chapman SJ, Videan EN, Evans E, Fritz J, Stoinski TS, Dixson AF, Gagneux P. 2007. Functional evidence for differences in sperm competition in humans and chimpanzees. Am J Phys Anthropol 134:274-280.

Andrade-Rocha FT. 2003. Semen analysis in laboratory practice: an overview of routine tests. J Clin Lab Anal 17:247-258.

Andres AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin SQ, Hurle B, Program NCS, Schwartzberg PL, Williamson SH, Bustamante CD et al. 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. PLoS Genet 6:e1001157.

Andres AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD et al. 2009. Targets of balancing selection in the human genome. Mol Biol Evol 26:2755-2764.

Aston KI, Carrell DT. 2009. Genome-wide study of single-nucleotide polymorphisms associated with azoospermia and severe oligozoospermia. J Androl 30:711-725.

Aston KI, Krausz C, Laface I, Ruiz-Castane E, Carrell DT. 2010. Evaluation of 172 candidate polymorphisms for association with oligozoospermia or azoospermia in a large cohort of men of European descent. Hum Reprod 25:1383-1397.

Babiarz-Magee L, Chen N, Seiberg M, Lin CB. 2004. The expression and activation of protease-activated receptor-2 correlate with skin color. Pigment Cell Res 17:241-251.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11:17-30.

Bartlett JD, Simmer JP. 1999. Proteinases in developing dental enamel. Crit Rev Oral Biol Med 10:425-441.

Basu Mallick C, Iliescu FM, Mols M, Hill S, Tamang R, Chaubey G, Goto R, Ho SY, Gallego Romero I, Crivellaro F et al. 2013. The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. PLoS Genet 9:e1003912.

Batruch I, Lecker I, Kagedan D, Smith CR, Mullen BJ, Grober E, Lo KC, Diamandis EP, Jarvi KA. 2011. Proteomic analysis of seminal plasma from normal volunteers and post-vasectomy patients identifies over 2000 proteins and candidate biomarkers of the urogenital system. J Proteome Res 10:941-953.

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R et al. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res 24:14-24.

Becker-Pauly C, Howel M, Walker T, Vlad A, Aufenvenne K, Oji V, Lottaz D, Sterchi EE, Debela M, Magdolen V et al. 2007. The alpha and beta subunits of the metalloprotease meprin are expressed in separate layers of human epidermis, revealing different functions in keratinocyte proliferation and differentiation. J Invest Dermatol 127:1115-1125.

Bernett MJ, Blaber SI, Scarisbrick IA, Dhanarajan P, Thompson SM, Blaber M. 2002. Crystal structure and biochemical characterization of human kallikrein 6 reveals that a trypsin-like kallikrein is expressed in the central nervous system. J Biol Chem 277:24562-24570.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111-1120.

Bhoola KD, Figueroa CD, Worthy K. 1992. Bioregulation of kinins: kallikreins, kininogens, and kininases. Pharmacol Rev 44:1-80.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3:201-212.

Bitoun E, Chavanas S, Irvine AD, Lonie L, Bodemer C, Paradisi M, Hamel-Teillac D, Ansai S, Mitsuhashi Y, Taieb A et al. 2002. Netherton syndrome: disease expression and spectrum of SPINK5 mutations in 21 families. J Invest Dermatol 118:352-361.

Borgono CA, Diamandis EP. 2004. The emerging roles of human tissue kallikreins in cancer. Nat Rev Cancer 4:876-890.

Borgono CA, Michael IP, Komatsu N, Jayakumar A, Kapadia R, Clayman GL, Sotiropoulou G, Diamandis EP. 2007. A potential role for multiple tissue kallikrein serine proteases in epidermal desquamation. J Biol Chem 282:3640-3652.

Brattsand M, Stefansson K, Lundh C, Haasum Y, Egelrud T. 2005. A proteolytic cascade of kallikreins in the stratum corneum. J Invest Dermatol 124:198-203.

Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, Ho KM, Ring S, Hurles M, Deloukas P et al. 2014. Cis and trans effects of human genomic variants on gene expression. PLoS Genet 10:e1004461.

Buddenkotte J, Stroh C, Engels IH, Moormann C, Shpacovitch VM, Seeliger S, Vergnolle N, Vestweber D, Luger TA, Schulze-Osthoff K et al. 2005. Agonists of proteinase-activated receptor-2 stimulate upregulation of intercellular cell adhesion molecule-1 in primary human keratinocytes via activation of NF-kappa B. J Invest Dermatol 124:38-45.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153-1157.

Bustamante CD, Ramachandran S. 2009. Evaluating signatures of sex-specific processes in the human genome. Nat Genet 41:8-10.

Cagliani R, Sironi M. 2013. Pathogen-driven selection in the human genome. Int J Evol Biol 2013:204240.

Carnahan-Craig SJ, Jensen-Seaman MI. 2014. Rates of evolution of hominoid seminal proteins are correlated with function and expression, rather than mating system. J Mol Evol 78:87-99.

Carnahan SJ, Jensen-Seaman MI. 2008. Hominoid seminal protein evolution and ancestral mating behavior. Am J Primatol 70:939-948.

Carrell DT, Aston KI. 2011. The search for SNPs, CNVs, and epigenetic variants associated with the complex disease of male infertility. Syst Biol Reprod Med 57:17-26.

Caubet C, Jonca N, Brattsand M, Guerrin M, Bernard D, Schmidt R, Egelrud T, Simon M, Serre G. 2004. Degradation of corneodesmosome proteins by two serine proteases of the kallikrein family, SCTE/KLK5/hK5 and SCCE/KLK7/hK7. J Invest Dermatol 122:1235-1244.

Cedenho AP. 2007. Evaluation of the subfertile male. Pp. 115-140. Male Infertility.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64.

Chavanas S, Bodemer C, Rochat A, Hamel-Teillac D, Ali M, Irvine AD, Bonafe JL, Wilkinson J, Taieb A, Barrandon Y et al. 2000. Mutations in SPINK5, encoding a serine protease inhibitor, cause Netherton syndrome. Nat Genet 25:141-142.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. Genome Res 20:393-402.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. PLoS Genet 1:e35.

Claw KG. 2013. Proteomic identification and evolutionary analysis of primate reproductive proteins. University of Washington.

Clements J, Hooper J, Dong Y, Harvey T. 2001. The expanded human kallikrein (KLK) gene family: genomic organisation, tissue-specific expression and potential functions. Biol Chem 382:5-14.

Clements JA, Willemsen NM, Myers SA, Dong Y. 2004. The Tissue Kallikrein Family of Serine Proteases: Functional Roles in Human Disease and Potential as Clinical Biomarkers. Crit Rev Clin Lab Sci 41:265-312.

Coelho M, Luiselli D, Bertorelle G, Lopes AI, Seixas S, Destro-Bisol G, Rocha J. 2005. Microsatellite variation and evolution of human lactase persistence. Hum Genet 117:329-339.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The role of geography in human adaptation. PLoS Genet 5:e1000500.

Corbo RM, Ulizzi L, Piombo L, Scacchi R. 2008. Study on a possible effect of four longevity candidate genes (ACE, PON1, PPAR-gamma, and APOE) on human fertility. Biogerontology 9:317-323.

Crespi BJ. 2010. The origins and evolution of genetic disease risk in modern humans. Ann N Y Acad Sci 1206:80-109.

Crisci JL, Jensen JD. 2012. Evolution of the Human Genome: Adaptive Changes. eLS. John Wiley & Sons, Ltd.

Darwin C. 1859. On The Origin Of Species By Means Of Natural Selection, Or The Preservation Of Favoured Races In The Struggle For Life.

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L. 2013. Evidence for polygenic adaptation to pathogens in the human genome. Mol Biol Evol 30:1544-1558.

de Lamirande E, Yoshida K, Yoshiike TM, Iwamoto T, Gagnon C. 2001. Semenogelin, the main protein of semen coagulum, inhibits human sperm capacitation by interfering with the superoxide anion generated during this process. J Androl 22:672-679.

Debela M, Beaufort N, Magdolen V, Schechter NM, Craik CS, Schmitt M, Bode W, Goettig P. 2008. Structures and specificity of the human kallikrein-related peptidases KLK 4, 5, 6, and 7. Biol Chem 389:623-632.

Debela M, Goettig P, Magdolen V, Huber R, Schechter NM, Bode W. 2007a. Structural basis of the zinc inhibition of human tissue kallikrein 5. J Mol Biol 373:1017-1031.

Debela M, Hess P, Magdolen V, Schechter NM, Steiner T, Huber R, Bode W, Goettig P. 2007b. Chymotryptic specificity determinants in the 1.0 A structure of the zinc-inhibited human tissue kallikrein 7. Proc Natl Acad Sci U S A 104:16086-16091.

Debela M, Magdolen V, Grimminger V, Sommerhoff C, Messerschmidt A, Huber R, Friedrich R, Bode W, Goettig P. 2006. Crystal structures of human tissue kallikrein 4: activity modulation by a specific zinc binding site. J Mol Biol 362:1094-1107.

Deperthes D, Frenette G, Brillard-Bourdet M, Bourgeois L, Gauthier F, Tremblay RR, Dube JY. 1996. Potential involvement of kallikrein hK2 in the hydrolysis of the human seminal vesicle proteins after ejaculation. J Androl 17:659-665.

Deraison C, Bonnart C, Lopez F, Besson C, Robinson R, Jayakumar A, Wagberg F, Brattsand M, Hachem JP, Leonardsson G et al. 2007. LEKTI fragments specifically inhibit KLK5, KLK7, and KLK14 and control desquamation through a pH-dependent interaction. Mol Biol Cell 18:3607-3619.

Descargues P, Deraison C, Bonnart C, Kreft M, Kishibe M, Ishida-Yamamoto A, Elias P, Barrandon Y, Zambruno G, Sonnenberg A et al. 2005. Spink5-deficient mice mimic Netherton syndrome through degradation of desmoglein 1 by epidermal protease hyperactivity. Nat Genet 37:56-65.

Descargues P, Deraison C, Prost C, Fraitag S, Mazereeuw-Hautier J, D'Alessio M, Ishida-Yamamoto A, Bodemer C, Zambruno G, Hovnanian A. 2006. Corneodesmosomal cadherins are preferential targets of stratum corneum trypsin- and chymotrypsin-like hyperactivity in Netherton syndrome. J Invest Dermatol 126:1622-1632.

Di Rienzo A. 2006. Population genetics models of common diseases. Curr Opin Genet Dev 16:630-636.

Di Rienzo A, Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21:596-601.

Diamandis EP, Scorilas A, Kishi T, Blennow K, Luo LY, Soosaipillai A, Rademaker AW, Sjogren M. 2004. Altered kallikrein 7 and 10 concentrations in cerebrospinal fluid of patients with Alzheimer's disease and frontotemporal dementia. Clin Biochem 37:230-237.

Ding X, Zhang J, Bian Z, Xia Y, Lu C, Gu A, Li Y, Song L, Wang S, Wang X. 2010a. Variants in the Eppin gene show association with semen quality in Han-Chinese population. Reprod Biomed Online 20:125-131.

Ding X, Zhang J, Fei J, Bian Z, Li Y, Xia Y, Lu C, Song L, Wang S, Wang X. 2010b. Variants of the EPPIN gene affect the risk of idiopathic male infertility in the Han-Chinese population. Hum Reprod 25:1657-1665.

Dixson AF. 2009. Sexual Selection and the Origins of Human Mating Systems. Oxford University Press Inc., New York, United States.

Dixson AL, Anderson MJ. 2002. Sexual selection, seminal coagulation and copulatory plug formation in primates. Folia Primatol (Basel) 73:63-69.

do Espirito Santo AR, Line SR. 2005. The enamel organic matrix: structure and function. Braz J Oral Sci 4:716-724.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat Genet 36:1326-1329.

Dorus S, Wasbrough ER, Busby J, Wilkin EC, Karr TL. 2010. Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. Mol Biol Evol 27:1235-1246.

Du Plessis SS, Gokul S, Agarwal A. 2013. Semen hyperviscosity: causes, consequences, and cures. Front Biosci (Elite Ed) 5:224-231.

Durand E, Tenaillon MI, Ridel C, Coubriche D, Jamin P, Jouanne S, Ressayre A, Charcosset A, Dillmann C. 2010. Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds. BMC Evol Biol 10:2.

Edstrom AM, Malm J, Frohm B, Martellini JA, Giwercman A, Morgelin M, Cole AM, Sorensen OE. 2008. The major bactericidal activity of human seminal plasma is zinc-dependent and derived from fragmentation of the semenogelins. J Immunol 181:3413-3421.

Egelrud T, Brattsand M, Kreutzmann P, Walden M, Vitzithum K, Marx UC, Forssmann WG, Magert HJ. 2005. hK5 and hK7, two serine proteinases abundant in human skin, are inhibited by LEKTI domain 6. Br J Dermatol 153:1200-1203.

Eirin-Lopez JM, Rebordinos L, Rooney AP, Rozas J. 2012. The birth-and-death evolution of multigene families revisited. Genome Dyn 7:170-196.

Eissa A, Diamandis EP. 2008. Human tissue kallikreins as promiscuous modulators of homeostatic skin barrier functions. Biol Chem 389:669-680.

Elliott MB, Irwin DM, Diamandis EP. 2006. In silico identification and Bayesian phylogenetic analysis of multiple new mammalian kallikrein gene families. Genomics 88:591-599.

Emami N, Deperthes D, Malm J, Diamandis EP. 2008. Major role of human KLK14 in seminal clot liquefaction. J Biol Chem 283:19561-19569.

Emami N, Diamandis EP. 2007. New insights into the functional mechanisms and clinical applications of the kallikrein-related peptidase family. Mol Oncol 1:269-287.

Emami N, Diamandis EP. 2008. Human kallikrein-related peptidase 14 (KLK14) is a new activator component of the KLK proteolytic cascade. Possible function in seminal plasma and skin. J Biol Chem 283:3031-3041.

Emami N, Scorilas A, Soosaipillai A, Earle T, Mullen B, Diamandis EP. 2009. Association between kallikrein-related peptidases (KLKs) and macroscopic indicators of semen analysis: their relation to sperm motility. Biol Chem 390:921-929.

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869-872.

Epstein DJ. 2009. Cis-regulatory mutations in human disease. Brief Funct Genomic Proteomic 8:310-316.

Excoffier L. 2002. Human demographic history: refining the recent African origin model. Curr Opin Genet Dev 12:675-682.

Excoffier L, Hofer T, Foll M. 2009. Detecting loci under selection in a hierarchically structured population. Heredity (Edinb) 103:285-298.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405-1413.

Ferreira Z, Hurle B, Andres AM, Kretzschmar WW, Mullikin JC, Cherukuri PF, Cruz P, Gonder MK, Stone AC, Tishkoff S et al. 2013a. Sequence diversity of Pan troglodytes subspecies and the impact of WFDC6 selective constraints in reproductive immunity. Genome Biol Evol 5:2512-2523.

Ferreira Z, Hurle B, Rocha J, Seixas S. 2011. Differing evolutionary histories of WFDC8 (short-term balancing) in Europeans and SPINT4 (incomplete selective sweep) in Africans. Mol Biol Evol 28:2811-2822.

Ferreira Z, Seixas S, Andres AM, Kretzschmar WW, Mullikin JC, Cherukuri PF, Cruz P, Swanson WJ, Program NCS, Clark AG et al. 2013b. Reproduction and immunity-driven natural selection in the human WFDC locus. Mol Biol Evol 30:938-950.

Findlay GD, Swanson WJ. 2010. Proteomics enhances evolutionary and functional analysis of reproductive proteins. Bioessays 32:26-36.

Fischer J, Meyer-Hoffert U. 2013. Regulation of kallikrein-related peptidases in the skin - from physiology to diseases to therapeutic options. Thromb Haemost 110:442-449.

Fortelny N, Cox JH, Kappelhoff R, Starr AE, Lange PF, Pavlidis P, Overall CM. 2014. Network analyses reveal pervasive functional regulation between proteases in the human protease web. PLoS Biol 12:e1001869.

Fortugno P, Bresciani A, Paolini C, Pazzagli C, El Hachem M, D'Alessio M, Zambruno G. 2011. Proteolytic Activation Cascade of the Netherton Syndrome-Defective Protein, LEKTI, in the Epidermis: Implications for Skin Homeostasis. J Invest Dermatol 131:2223-2232.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915-925.

Fuhrman-Luck RA, Silva ML, Dong Y, Irving-Rodgers H, Stoll T, Hastie ML, Loessner D, Gorman JJ, Clements JA. 2014. Proteomic and other analyses to determine the functional consequences of deregulated kallikrein-related peptidase (KLK) expression in prostate and ovarian cancer. PROTEOMICS – Clinical Applications 8:403-415.

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T et al. 2008. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. Hum Mol Genet 17:835-843.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res 19:199-212.

Fumagalli M, Sironi M. 2014. Human genome variability, natural selection and infectious diseases. Curr Opin Immunol 30C:9-16.

Furio L, de Veer S, Jaillet M, Briot A, Robin A, Deraison C, Hovnanian A. 2014. Transgenic kallikrein 5 mice reproduce major cutaneous and systemic hallmarks of Netherton syndrome. J Exp Med 211:499-513.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M et al. 2002. The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Gan L, Lee I, Smith R, Argonza-Barrett R, Lei H, McCuaig J, Moss P, Paeper B, Wang K. 2000. Sequencing and expression analysis of the serine protease gene cluster located in chromosome 19q13 region. Gene 257:119-130.

Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

Genomes Project C, Corresponding a, Steering c, Production g, Baylor College of M, Shenzhen BGI, Broad Institute of MIT, Harvard, Coriell Institute for Medical R, European Molecular Biology Laboratory EBI et al. 2015. A global reference for human genetic variation. Nature 526:68-74.

Gomes S, Marques PI, Matthiesen R, Seixas S. 2014. Adaptive evolution and divergence of SERPINB3: a young duplicate in great Apes. PLoS One 9:e104935.

Gomis-Ruth FX, Bayes A, Sotiropoulou G, Pampalakis G, Tsetsenis T, Villegas V, Aviles FX, Coll M. 2002. The structure of human prokallikrein 6 reveals a novel activation mechanism for the kallikrein family. J Biol Chem 277:27273-27281.

Good JM, Wiebe V, Albert FW, Burbano HA, Kircher M, Green RE, Halbwax M, Andre C, Atencia R, Fischer A et al. 2013. Comparative population genomics of the ejaculate in humans and the great apes. Mol Biol Evol 30:964-976.

Goudet G, Mugnier S, Callebaut I, Monget P. 2008. Phylogenetic analysis and identification of pseudogenes reveal a progressive loss of zona pellucida genes during evolution of vertebrates. Biol Reprod 78:796-806.

Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. 2012. Limited evidence for classic selective sweeps in African populations. Genetics 192:1049-1064.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Genomes P, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108:11983-11988.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH et al. 2010. A draft sequence of the Neandertal genome. Science 328:710-722.

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell 152:703-713.

Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883-886.

GTEx-Consortium. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348:648-660.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695.

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A. 2010a. Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philos Trans R Soc Lond B Biol Sci 365:2459-2468.

Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, Coop G et al. 2010b. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc Natl Acad Sci U S A 107 Suppl 2:8924-8930.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet 4:e32.

Hanihara T. 2008. Morphological variation of major human populations based on nonmetric dental traits. Am J Phys Anthropol 136:169-182.

Hanihara T, Ishida H. 2005. Metric dental variation of major human populations. Am J Phys Anthropol 128:287-298.

Harcourt AH, Harvey PH, Larson SG, Short RV. 1981. Testis weight, body weight and breeding system in primates. Nature 293:55-57.

Harris K, Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet 9:e1003521.

Hart PS, Hart TC, Michalec MD, Ryu OH, Simmons D, Hong S, Wright JT. 2004. Mutation in kallikrein 4 causes autosomal recessive hypomaturation amelogenesis imperfecta. J Med Genet 41:545-549.

Helgason A, Palsson S, Thorleifsson G, Grant SF, Emilsson V, Gunnarsdottir S, Adeyemo A, Chen Y, Chen G, Reynisdottir I et al. 2007. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat Genet 39:218-225.

Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335-2352.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Genomes P, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. Science 331:920-924.

Hu JC, Sun X, Zhang C, Liu S, Bartlett JD, Simmer JP. 2002. Enamelysin and kallikrein-4 mRNA expression in developing mouse molars. Eur J Oral Sci 110:307-315.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

Huff CD, Witherspoon DJ, Zhang Y, Gatenbee C, Denson LA, Kugathasan S, Hakonarson H, Whiting A, Davis CT, Wu W et al. 2012. Crohn's disease and genetic hitchhiking at IBD5. Mol Biol Evol 29:101-111.

Hurle B, Swanson W, Program NCS, Green ED. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome Res 17:276-286.

Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci U S A 101:10667-10672.

International HapMap C. 2005. A haplotype map of the human genome. Nature 437:1299-1320.

International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Ishida-Yamamoto A, Igawa S. 2014. The biology and regulation of corneodesmosomes. Cell and Tissue Research.

Ishida-Yamamoto A, Igawa S, Kishibe M. 2011. Order and disorder in corneocyte adhesion. J Dermatol 38:645-654.

Iwata T, Yamakoshi Y, Hu JC, Ishikawa I, Bartlett JD, Krebsbach PH, Simmer JP. 2007. Processing of ameloblastin by MMP-20. J Dent Res 86:153-157.

Jaffa AA, Chai KX, Chao J, Chao L, Mayfield RK. 1992. Effects of diabetes and insulin on expression of kallikrein and renin genes in the kidney. Kidney Int 41:789-795.

Jensen-Seaman MI, Li WH. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. J Mol Evol 57:261-270.

Jequier AM. 2000. The Anatomy and Physiology of the Male Genital Tract. Male Infertility: A Guide for the Clinician. Cambridge University Press.

Jha P, Lu D, Xu S. 2015. Natural Selection and Functional Potentials of Human Noncoding Elements Revealed by Analysis of Next Generation Sequencing Data. PLoS One 10:e0129023.

Jonsson M, Linse S, Frohm B, Lundwall A, Malm J. 2005. Semenogelins I and II bind zinc and regulate the activity of prostate-specific antigen. Biochem J 387:447-453.

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H et al. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152:691-702.

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. Nat Rev Genet 15:379-393.

Kaushal A, Myers SA, Dong Y, Lai J, Tan OL, Bui LT, Hunt ML, Digby MR, Samaratunga H, Gardiner RA et al. 2008. A novel transcript from the KLKP1 gene is androgen regulated, down-regulated during prostate cancer progression and encodes the first non-serine protease identified from the human kallikrein gene locus. Prostate 68:381-399.

Kawasaki K, Hu JC, Simmer JP. 2014. Evolution of Klk4 and enamel maturation in eutherians. Biol Chem 395:1003-1013.

Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PI et al. 2015. Insights into the evolution of longevity from the bowhead whale genome. Cell Rep 10:112-122.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336:740-743.

Key FM, Teixeira JC, de Filippo C, Andres AM. 2014. Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev 29:45-51.

Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, Haneji K, Hanihara T, Matsukusa H, Kawamura S, Maki K et al. 2009. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. Am J Hum Genet 85:528-535.

Kishibe M. 2014. Kallikrein-Related Peptidase 8 (KLK8): The Structure and Function in the Epidermis. Journal of Dermatology and Clinical Research 2.

Klokk TI, Xi Z, Saatcioglu F. 2006. Human Tissue Kallikreins - A Family with many surprises. Turkish Journal of Biochemistry 31:69-78.

Komatsu N, Saijoh K, Jayakumar A, Clayman GL, Tohyama M, Suga Y, Mizuno Y, Tsukamoto K, Taniuchi K, Takehara K et al. 2008. Correlation between SPINK5

gene mutations and clinical manifestations in Netherton syndrome patients. J Invest Dermatol 128:1148-1159.

Komatsu N, Saijoh K, Kuk C, Liu AC, Khan S, Shirasaki F, Takehara K, Diamandis EP. 2007a. Human tissue kallikrein expression in the stratum corneum and serum of atopic dermatitis patients. Exp Dermatol 16:513-519.

Komatsu N, Saijoh K, Kuk C, Shirasaki F, Takehara K, Diamandis EP. 2007b. Aberrant human tissue kallikrein levels in the stratum corneum and serum of patients with psoriasis: dependence on phenotype, severity and therapy. Br J Dermatol 156:875-883.

Komatsu N, Saijoh K, Toyama T, Ohka R, Otsuki N, Hussack G, Takehara K, Diamandis EP. 2005. Multiple tissue kallikrein mRNA and protein expression in normal skin and skin diseases. Br J Dermatol 153:274-281.

Komatsu N, Takata M, Otsuki N, Ohka R, Amano O, Takehara K, Saijoh K. 2002. Elevated stratum corneum hydrolytic activity in Netherton syndrome suggests an inhibitory regulation of desquamation by SPINK5-derived peptides. J Invest Dermatol 118:436-443.

Komatsu N, Takata M, Otsuki N, Toyama T, Ohka R, Takehara K, Saijoh K. 2003. Expression and localization of tissue kallikrein mRNAs in human epidermis and appendages. J Invest Dermatol 121:542-549.

Kontos CK, Scorilas A. 2012. Kallikrein-related peptidases (KLKs): a gene family of novel cancer biomarkers. Clin Chem Lab Med 50:1877-1891.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144.

Koumandou VL, Scorilas A. 2013. Evolution of the plasma and tissue kallikreins, and their alternative splicing isoforms. PLoS One 8:e68074.

Kurlender L, Borgono C, Michael IP, Obiezu C, Elliott MB, Yousef GM, Diamandis EP. 2005. A survey of alternative transcripts of human tissue kallikrein genes. Biochim Biophys Acta 1755:1-14.

Laflamme BA, Wolfner MF. 2013. Identification and function of proteolysis regulators in seminal fluid. Mol Reprod Dev 80:80-101.

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310:1782-1786.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832-838.

Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. PLoS One 5:e10284.

Lawrence MG, Lai J, Clements JA. 2010. Kallikreins on Steroids: Structure, Function, and Hormonal Regulation of Prostate-Specific Antigen and the Extended Kallikrein Locus. Endocr Rev 31:407-446.

Laxmikanthan G, Blaber SI, Bernett MJ, Scarisbrick IA, Juliano MA, Blaber M. 2005. 1.70 A X-ray structure of human apo kallikrein 1: structural changes upon peptide inhibitor/substrate binding. Proteins 58:802-814.

Lee SE, Jeong SK, Lee SH. 2010. Protease and protease-activated receptor-2 signaling in the pathogenesis of atopic dermatitis. Yonsei Med J 51:808-822.

Lettre G. 2014. Rare and low-frequency variants in human common diseases and other complex traits. J Med Genet 51:705-714.

Lilja H. 1985. A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. J Clin Invest 76:1899-1903.

Lilja H, Abrahamsson PA, Lundwall A. 1989. Semenogelin, the predominant protein in human semen. Primary structure and identification of closely related proteins in the male accessory sex glands and on the spermatozoa. J Biol Chem 264:1894-1900.

Lilja H, Laurell CB. 1985. The predominant protein in human seminal coagulate. Scand J Clin Lab Invest 45:635-641.

Lilja H, Oldbring J, Rannevik G, Laurell CB. 1987. Seminal vesicle-secreted proteins and their reactions during gelation and liquefaction of human semen. J Clin Invest 80:281-285.

Lopez-Otin C, Bond JS. 2008. Proteases: multifunctional enzymes in life and disease. J Biol Chem 283:30433-30437.

Lu W, Zhou D, Glusman G, Utleg AG, White JT, Nelson PS, Vasicek TJ, Hood L, Lin B. 2006. KLK31P is a novel androgen regulated and transcribed pseudogene of kallikreins that is expressed at lower levels in prostate cancer cells than in normal prostate cells. Prostate 66:936-944.

Lu Y, Papagerakis P, Yamakoshi Y, Hu JC, Bartlett JD, Simmer JP. 2008. Functions of KLK4 and MMP-20 in dental enamel formation. Biol Chem 389:695-700.

Luca F, Perry GH, Di Rienzo A. 2010. Evolutionary adaptations to dietary changes. Annu Rev Nutr 30:291-314.

Lundwall A. 2013. Old genes and new genes: the evolution of the kallikrein locus. Thromb Haemost 110:469-475.

Lundwall A, Band V, Blaber M, Clements JA, Courty Y, Diamandis EP, Fritz H, Lilja H, Malm J, Maltais LJ et al. 2006a. A comprehensive nomenclature for serine proteases with homology to tissue kallikreins. Biol Chem 387:637-641.

Lundwall A, Brattsand M. 2008. Kallikrein-related peptidases. Cell Mol Life Sci 65:2019-2038.

Lundwall A, Clauss A, Olsson AY. 2006b. Evolution of kallikrein-related peptidases in mammals and identification of a genetic locus encoding potential regulatory inhibitors. Biol Chem 387:243-249.

Malm J, Hellman J, Magnusson H, Laurell CB, Lilja H. 1996. Isolation and characterization of the major gel proteins in human semen, semenogelin I and semenogelin II. Eur J Biochem 238:48-53.

Malm J, Jonsson M, Frohm B, Linse S. 2007. Structural properties of semenogelin I. FEBS J 274:4503-4510.

Martinez-Heredia J, de Mateo S, Vidal-Taboada JM, Ballesca JL, Oliva R. 2008. Identification of proteomic differences in asthenozoospermic sperm samples. Hum Reprod 23:783-791.

Matsumura M, Bhatt AS, Andress D, Clegg N, Takayama TK, Craik CS, Nelson PS. 2005. Substrates of the prostate-specific serine protease prostase/KLK4 defined by positional-scanning peptide libraries. Prostate 62:1-13.

McCrudden MT, Dafforn TR, Houston DF, Turkington PT, Timson DJ. 2008. Functional domains of the human epididymal protease inhibitor, eppin. FEBS J 275:1742-1750.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

McEvoy B, Beleza S, Shriver MD. 2006. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. Hum Mol Genet 15 Spec No 2:R176-181.

Memari N, Jiang W, Diamandis EP, Luo LY. 2007. Enzymatic properties of human kallikrein-related peptidase 12 (KLK12). Biol Chem 388:427-435.

Menez R, Michel S, Muller BH, Bossus M, Ducancel F, Jolivet-Reynaud C, Stura EA. 2008. Crystal structure of a ternary complex between human prostate-specific

antigen, its substrate acyl intermediate and an activating antibody. J Mol Biol 376:1021-1033.

Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28:659-669.

Michael IP, Pampalakis G, Mikolajczyk SD, Malm J, Sotiropoulou G, Diamandis EP. 2006. Human tissue kallikrein 5 is a member of a proteolytic cascade pathway involved in seminal clot liquefaction and potentially in prostate cancer progression. J Biol Chem 281:12743-12750.

Michael IP, Sotiropoulou G, Pampalakis G, Magklara A, Ghosh M, Wasney G, Diamandis EP. 2005. Biochemical and enzymatic characterization of human kallikrein 5 (hK5), a novel serine protease potentially involved in cancer progression. J Biol Chem 280:14628-14635.

Mubiru JN, Yang AS, Olsen C, Nayak S, Livi CB, Dick EJ, Jr., Owston M, Garcia-Forey M, Shade RE, Rogers J. 2014. Analysis of prostate-specific antigen transcripts in chimpanzees, cynomolgus monkeys, baboons, and African green monkeys. PLoS One 9:e94522.

Murakami K, Jiang YP, Tanaka T, Bando Y, Mitrovic B, Yoshida S. 2013. In vivo analysis of kallikrein-related peptidase 6 (KLK6) function in oligodendrocyte development and the expression of myelin proteins. Neuroscience 236:1-11.

Nagano T, Kakegawa A, Yamakoshi Y, Tsuchiya S, Hu JC, Gomi K, Arai T, Bartlett JD, Simmer JP. 2009. Mmp-20 and Klk4 cleavage site preferences for amelogenin sequences. J Dent Res 88:823-828.

Neel JV. 1962. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14:353-362.

Neel JV, Weder AB, Julius S. 1998. Type II diabetes, essential hypertension, and obesity as "syndromes of impaired genetic homeostasis": the "thrifty genotype" hypothesis enters the 21st century. Perspect Biol Med 42:44-74.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121-152.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6:e1000888.

Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet 39:197-218.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3:e170.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Nishifuji K, Yoon JS. 2013. The stratum corneum: the rampart of the mammalian body. Vet Dermatol 24:60-72 e15-66.

Ny A, Egelrud T. 2004. Epidermal hyperproliferation and decreased skin barrier function in mice overexpressing stratum corneum chymotryptic enzyme. Acta Derm Venereol 84:18-22.

Obiezu CV, Diamandis EP. 2005. Human tissue kallikrein gene family: applications in cancer. Cancer Lett 224:1-22.

Ogawa K, Yamada T, Tsujioka Y, Taguchi J, Takahashi M, Tsuboi Y, Fujino Y, Nakajima M, Yamamoto T, Akatsu H et al. 2000. Localization of a novel type trypsin-like serine protease, neurosin, in brain tissues of Alzheimer's disease and Parkinson's disease. Psychiatry Clin Neurosci 54:419-426.

Ohno S. 1970. Evolution by Gene Duplication. Springer.

Oka T, Hakoshima T, Itakura M, Yamamori S, Takahashi M, Hashimoto Y, Shiosaka S, Kato K. 2002. Role of loop structures of neuropsin in the activity of serine protease and regulated secretion. J Biol Chem 277:14724-14730.

Olsson AY, Valtonen-Andre C, Lilja H, Lundwall A. 2004. The evolution of the glandular kallikrein locus: identification of orthologs and pseudogenes in the cotton-top tamarin. Gene 343:347-355.

Organization WH. 1992. Recent advances in medically assisted conception. Report of a WHO Scientific Group. World Health Organ Tech Rep Ser 820:1-111.

Ovaere P, Lippens S, Vandenabeele P, Declercq W. 2009. The emerging roles of serine protease cascades in the epidermis. Trends Biochem Sci 34:453-463.

Pavlopoulou A, Pampalakis G, Michalopoulos I, Sotiropoulou G. 2010. Evolutionary history of tissue kallikreins. PLoS One 5:e13781.

Pennings PS, Hermisson J. 2006a. Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23:1076-1084.

Pennings PS, Hermisson J. 2006b. Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet 2:e186.

Peter A, Lilja H, Lundwall A, Malm J. 1998. Semenogelin I and semenogelin II, the major gel-forming proteins in human semen, are substrates for transglutaminase. Eur J Biochem 252:216-221.

Peter BM, Huerta-Sanchez E, Nielsen R. 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet 8:e1003011.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19:826-837.

Pilch B, Mann M. 2006. Large-scale and high-confidence proteomic analysis of human seminal plasma. Genome Biol 7:R40.

Pons S, Griol-Charhbili V, Heymes C, Fornes P, Heudes D, Hagege A, Loyer X, Meneton P, Giudicelli JF, Samuel JL et al. 2008. Tissue kallikrein deficiency aggravates cardiac remodelling and decreases survival after myocardial infarction in mice. Eur J Heart Fail 10:343-351.

Practice Committee of American Society for Reproductive M. 2012. Diagnostic evaluation of the infertile male: a committee opinion. Fertil Steril 98:294-301.

Prassas I, Eissa A, Poda G, Diamandis EP. 2015. Unleashing the therapeutic potential of human kallikrein-related serine proteases. Nat Rev Drug Discov 14:183-202.

Pritchard JK, Di Rienzo A. 2010. Adaptation - not by sweeps alone. Nat Rev Genet 11:665-667.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20:R208-215.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. Evolution 59:2312-2323.

Puente XS, López-Otín C. 2004. A Genomic Analysis of Rat Proteases and Protease Inhibitors. Genome Res 14:609-622.

Puente XS, Sanchez LM, Overall CM, Lopez-Otin C. 2003. Human and mouse proteases: a comparative genomic approach. Nat Rev Genet 4:544-558.

Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreno-Torres A, Pavlidis P, Laayouni H, Bertranpetit J, Engelken J. 2014. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Res 42:D903-909.

Quesada V, Ordonez GR, Sanchez LM, Puente XS, Lopez-Otin C. 2009. The Degradome database: mammalian proteases and diseases of proteolysis. Nucleic Acids Res 37:D239-243.

Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 23:437-441.

Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. Mol Biol Evol 25:207-219.

Ramsay AJ, Dong Y, Hunt ML, Linn M, Samaratunga H, Clements JA, Hooper JD. 2008. Kallikrein-related peptidase 4 (KLK4) initiates intracellular signaling via protease-

activated receptors (PARs). KLK4 and PAR-2 are co-expressed during prostate cancer progression. J Biol Chem 283:12293-12304.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053-1060.

Riegman PH, Vlietstra RJ, Suurmeijer L, Cleutjens CB, Trapman J. 1992. Characterization of the human kallikrein locus. Genomics 14:6-11.

Robert M, Gagnon C. 1999. Semenogelin I: a coagulum forming, multifunctional seminal vesicle protein. Cell Mol Life Sci 55:944-960.

Ryu O, Hu JC, Yamakoshi Y, Villemain JL, Cao X, Zhang C, Bartlett JD, Simmer JP. 2002. Porcine kallikrein-4 activation, glycosylation, activity, and expression in prokaryotic and eukaryotic hosts. Eur J Oral Sci 110:358-365.

Ryu OH, Fincham AG, Hu CC, Zhang C, Qian Q, Bartlett JD, Simmer JP. 1999. Characterization of recombinant pig enamelysin activity and cleavage of recombinant pig and mouse amelogenins. J Dent Res 78:743-750.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312:1614-1620.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918.

Scheinfeldt LB, Tishkoff SA. 2013. Recent human adaptation: genomic approaches, interpretation and insights. Nat Rev Genet 14:692-702.

Seiberg M, Paine C, Sharlow E, Andrade-Gordon P, Costanzo M, Eisinger M, Shapiro SS. 2000. The protease-activated receptor 2 regulates pigmentation via keratinocyte-melanocyte interactions. Exp Cell Res 254:25-32.

Seixas S, Ivanova N, Ferreira Z, Rocha J, Victor BL. 2012. Loss and gain of function in SERPINB11: an example of a gene under selection on standing variation, with implications for host-pathogen interactions. PLoS One 7:e32518.

Sharma JN. 2003. Does the kinin system mediate in cardiovascular abnormalities? An overview. J Clin Pharmacol 43:1187-1195.

Sharma R, Agarwal A, Mohanty G, Jesudasan R, Gopalan B, Willard B, Yadav SP, Sabanegh E. 2013. Functional proteomic analysis of seminal plasma proteins in men with various semen parameters. Reprod Biol Endocrinol 11:38.

Sharpe RM, Franks S. 2002. Environment, lifestyle and infertility--an inter-generational issue. Nat Cell Biol 4 Suppl:s33-40.

Shaw JL, Diamandis EP. 2007. Distribution of 15 human kallikreins in tissues and biological fluids. Clin Chem 53:1423-1432.

Shiina T, Ota M, Shimizu S, Katsuyama Y, Hashimoto N, Takasu M, Anzai T, Kulski JK, Kikkawa E, Naruse T et al. 2006. Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. Genetics 173:1555-1570.

Shimizu-Okabe C, Yousef GM, Diamandis EP, Yoshida S, Shiosaka S, Fahnestock M. 2001. Expression of the kallikrein gene family in normal and Alzheimer's disease brain. Neuroreport 12:2747-2751.

Simmer JP, Hu JC. 2002. Expression, structure, and function of enamel proteinases. Connect Tissue Res 43:441-449.

Simmer JP, Hu Y, Lertlam R, Yamakoshi Y, Hu JC. 2009. Hypomaturation enamel defects in Klk4 knockout/LacZ knockin mice. J Biol Chem 284:19110-19121.

Smith CE, Richardson AS, Hu Y, Bartlett JD, Hu JC, Simmer JP. 2011. Effect of kallikrein 4 loss on enamel mineralization: comparison with mice lacking matrix metalloproteinase 20. J Biol Chem 286:18149-18160.

Steinhoff M, Neisius U, Ikoma A, Fartasch M, Heyer G, Skov PS, Luger TA, Schmelz M. 2003. Proteinase-activated receptor-2 mediates itch: a novel pathway for pruritus in human skin. J Neurosci 23:6176-6180.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526:75-81.

Sun C, Southard C, Witonsky DB, Kittler R, Di Rienzo A. 2010. Allele-specific down-regulation of RPTOR expression induced by retinoids contributes to climate adaptations. PLoS Genet 6:e1001178.

Suzuki K, Kise H, Nishioka J, Hayashi T. 2007. The interaction among protein C inhibitor, prostate-specific antigen, and the semenogelin system. Semin Thromb Hemost 33:46-52.

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci U S A 98:2509-2514.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Takayama TK, Carter CA, Deng T. 2001a. Activation of prostate-specific antigen precursor (pro-PSA) by prostin, a novel human prostatic serine protease identified by degenerate PCR. Biochemistry 40:1679-1687.

Takayama TK, McMullen BA, Nelson PS, Matsumura M, Fujikawa K. 2001b. Characterization of hK4 (prostase), a prostate-specific serine protease: activation of the precursor of prostate specific antigen (pro-PSA) and single-chain urokinase-type plasminogen activator and degradation of prostatic acid phosphatase. Biochemistry 40:15341-15348.

Tamura H, Kawata M, Hamaguchi S, Ishikawa Y, Shiosaka S. 2012. Processing of neuregulin-1 by neuropsin regulates GABAergic neuron to control neural plasticity of the mouse hippocampus. J Neurosci 32:12657-12672.

Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, Nickel B, Fischer A, Halbwax M, Andre C et al. 2015. Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees and bonobos. Mol Biol Evol.

Tian X, Pascal G, Fouchecourt S, Pontarotti P, Monget P. 2009. Gene birth, death, and divergence: the different scenarios of reproduction-related gene evolution. Biol Reprod 80:616-621.

Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Genetic Investigation of ATC, Hirschhorn JN. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet 44:1015-1019.

Turner CG, 2nd. 1990. Major features of Sundadonty and Sinodonty, including suggestions about East Asian microevolution, population history, and late Pleistocene relationships with Australian aboriginals. Am J Phys Anthropol 82:295-317.

Tye CE, Pham CT, Simmer JP, Bartlett JD. 2009. DPPI may activate KLK4 during enamel formation. J Dent Res 88:323-327.

Vaisanen V, Lovgren J, Hellman J, Piironen T, Lilja H, Pettersson K. 1999. Characterization and processing of prostate specific antigen (hK3) and human glandular kallikrein (hK2) secreted by LNCaP cells. Prostate Cancer Prostatic Dis 2:91-97.

Vasseur E, Quintana-Murci L. 2013. The impact of natural selection on health and disease: uses of the population genetics approach in humans. Evol Appl 6:596-607.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. 2001. The sequence of the human genome. Science 291:1304-1351.

Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, Akey JM. 2012. Personal and population genomics of human regulatory variation. Genome Res 22:1689-1697.

Verrelli BC, Tishkoff SA, Stone AC, Touchman JW. 2006. Contrasting histories of G6PD molecular evolution and malarial resistance in humans and chimpanzees. Mol Biol Evol 23:1592-1601.

Vine MF, Margolin BH, Morrison HI, Hulka BS. 1994. Cigarette smoking and sperm density: a meta-analysis. Fertil Steril 61:35-43.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu Rev Genet 47:97-120.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A 102:18508-18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol 4:e72.

Walley AJ, Chavanas S, Moffatt MF, Esnouf RM, Ubhi B, Lawrence R, Wong K, Abecasis GR, Jones EY, Harper JI et al. 2001. Gene polymorphism in Netherton and common atopic disease. Nat Genet 29:175-178.

Wang SK, Hu Y, Simmer JP, Seymen F, Estrella NM, Pal S, Reid BM, Yildirim M, Bayram M, Bartlett JD et al. 2013. Novel KLK4 and MMP20 mutations discovered by whole-exome sequencing. J Dent Res 92:266-271.

Wang Z, Widgren EE, Richardson RT, O'Rand MG. 2007. Characterization of an eppin protein complex from human semen and spermatozoa. Biol Reprod 77:476-484.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet 3:e90.

Wright JT, Daly B, Simmons D, Hong S, Hart SP, Hart TC, Atsawasuwan P, Yamauchi M. 2006. Human enamel phenotype associated with amelogenesis imperfecta and a kallikrein-4 (g.2142G>A) proteinase mutation. Eur J Oral Sci 114 Suppl 1:13-17; discussion 39-41, 379.

Wright JT, Torain M, Long K, Seow K, Crawford P, Aldred MJ, Hart PS, Hart TC. 2011. Amelogenesis imperfecta: genotype-phenotype studies in 71 families. Cells Tissues Organs 194:279-283.

Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071-1076.

Yamakoshi Y, Hu JC, Fukae M, Yamakoshi F, Simmer JP. 2006. How do enamelysin and kallikrein 4 process the 32-kDa enamelin? Eur J Oral Sci 114 Suppl 1:45-51; discussion 93-45, 379-380.

Yamakoshi Y, Richardson AS, Nunez SM, Yamakoshi F, Milkovich RN, Hu JC, Bartlett JD, Simmer JP. 2011. Enamel proteins and proteases in Mmp20 and Klk4 null and double-null mice. Eur J Oral Sci 119 Suppl 1:206-216.

Yamakoshi Y, Simmer JP, Bartlett JD, Karakida T, Oida S. 2013. MMP20 and KLK4 activation and inactivation interactions in vitro. Arch Oral Biol 58:1569-1577.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908-917.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107-1118.

Yoon H, Blaber SI, Debela M, Goettig P, Scarisbrick IA, Blaber M. 2009. A completed KLK activome profile: investigation of activation profiles of KLK9, 10, and 15. Biol Chem 390:373-377.

Yoon H, Laxmikanthan G, Lee J, Blaber SI, Rodriguez A, Kogot JM, Scarisbrick IA, Blaber M. 2007. Activation profiles and regulatory cascades of the human kallikrein-related peptidases. J Biol Chem 282:31852-31864.

Yousef GM, Borgono CA, Michael IP, Diamandis EP. 2004. Cloning of a kallikrein pseudogene. Clin Biochem 37:961-967.

Yousef GM, Chang A, Scorilas A, Diamandis EP. 2000. Genomic organization of the human kallikrein gene family on chromosome 19q13.3-q13.4. Biochem Biophys Res Commun 276:125-133.

Yousef GM, Diamandis EP. 2001. The New Human Tissue Kallikrein Gene Family: Structure, Function, and Association to Disease. Endocr Rev 22:184-204.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174:1431-1439.

Zhang J, Webb DM, Podlaha O. 2002. Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. Genetics 162:1825-1835.

Zhao H, Lee WH, Shen JH, Li H, Zhang Y. 2008. Identification of novel semenogelin I-derived antimicrobial peptide from liquefied human seminal plasma. Peptides 29:505-511.

Zhu L, Liu H, Witkowska HE, Huang Y, Tanimoto K, Li W. 2014. Preferential and selective degradation and removal of amelogenin adsorbed on hydroxyapatites by MMP20 and KLK4 in vitro. Front Physiol 5.

Appendices

Appendix A - Supplementary Material Paper I

Birth-and-Death of *KLK3* and *KLK2* in Primates: Evolution Driven by Reproductive Biology

Genome Biol. Evol. 2012 4(12):1331-1338

Table S1 – General features of comparative sequence data set.

Species	N[a]	Sequence	Origin
Homo sapiens (Hsa)	1	GRCh37 NM_005551 (*KLK2*)	
	1	GRCh37 NM_001648 (*KLK3*)	
Pan troglodytes (Ptr)	1	CGSC 2.1.3/panTro3 (chr19:55720749-56012427)	Yerkes Primate Research Center in Atlanta, GA, USA
	2	ECACC[b] cell line EB176 (JC)	
Pan paniscus (Ppa)	1	AJFE01043435.1	
Gorilla gorilla (Ggo)	1	gorGor3.1/ gorGor3 (chr19:48160184-48948382)	Western Lowland
	2	ECACC[b] cell line EB (JC)	
	4	Our study	Lisbon Zoo
Pongo pygmaeus (Ppy)	1	WUGSC 2.0.2/ponAbe2 (chr19:52377104-52721866)	Gladys Porter Zoo in Brownsville
	2	ECACC[b] cell line EB185 (JC)	
	1	AC214871	BAC PAC resources bacpac.chori.org
Macaca mulatta (Mmu)	1	MGSC Merged 1.0/rheMac2 (chr19:57001913-57290885)	
Macaca fascicularis (Mfa)	2	ECACC[b] cell line CYNOM-K1	
Macaca fuscata (Mfu)	30	Our study	Lisbon Zoo
Saimiri boliviensis (Sbo)	1	AC199261	BAC PAC resources bacpac.chori.org
Saguinus oedipus (Soe)	1	AY556462	
Callithrix jacchus (Cja)	1	WUGSC 3.2/calJac3 (chr22:42634463-43227966)	Southwestern National Primate Research Center in San Antonio, TX, USA
Nomascus leucogenys (Nle)	1	AC198555	BAC PAC resources bacpac.chori.org
Hylobates sp.	2	ECACC[b] cell line MLA144	
	8	Our study	Lisbon Zoo
Papio anubis (Pan)	1	AC157440	BAC PAC resources bacpac.chori.org
Chlorocebus aethiops (Cae)	1	AC207140	BAC PAC resources bacpac.chori.org
Colobus guereza (Cgu)	1	AC153315	BAC PAC resources bacpac.chori.org
Aotus nancymaae (Ana)	1	AC153312	BAC PAC resources bacpac.chori.org
Callicebus moloch (Cmo)	1	AC204813	BAC PAC resources bacpac.chori.org
Ateles geoffroyi (Age)	1	AC225682	BAC PAC resources bacpac.chori.org
Eulemur macaco (Ema)	1	AC198556	BAC PAC resources bacpac.chori.org
Lemur catta (Lca)	1	AC153325	BAC PAC resources bacpac.chori.org
Otolemur garnettii (Oga)	1	AC153738	BAC PAC resources bacpac.chori.org

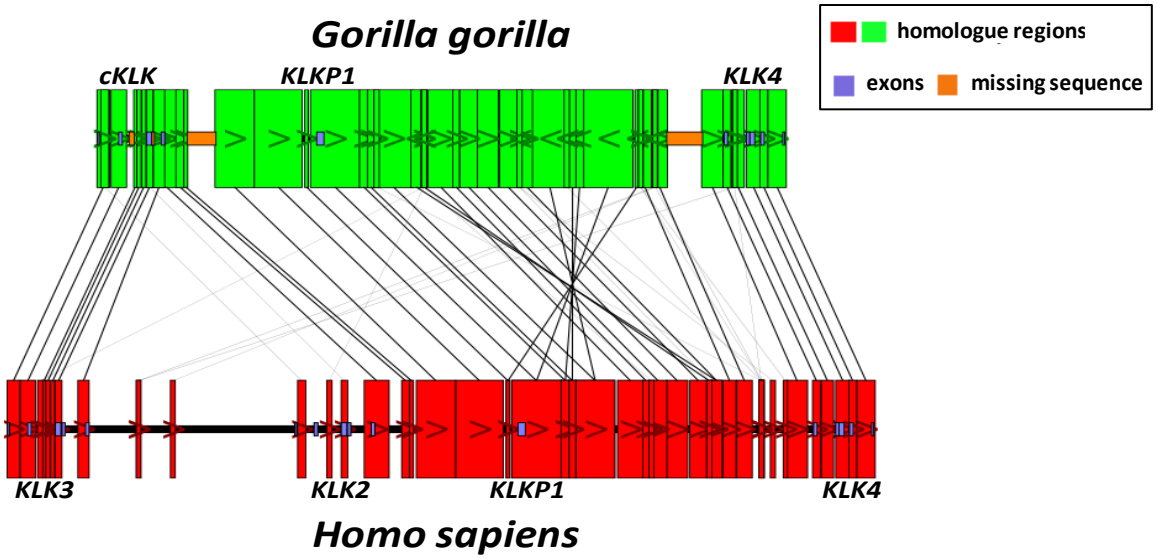
[a] N – sequence (chromosome) number
[b] European Collection of Cell Cultures

Table S2 – KLKs, SEMGs and sperm competition.

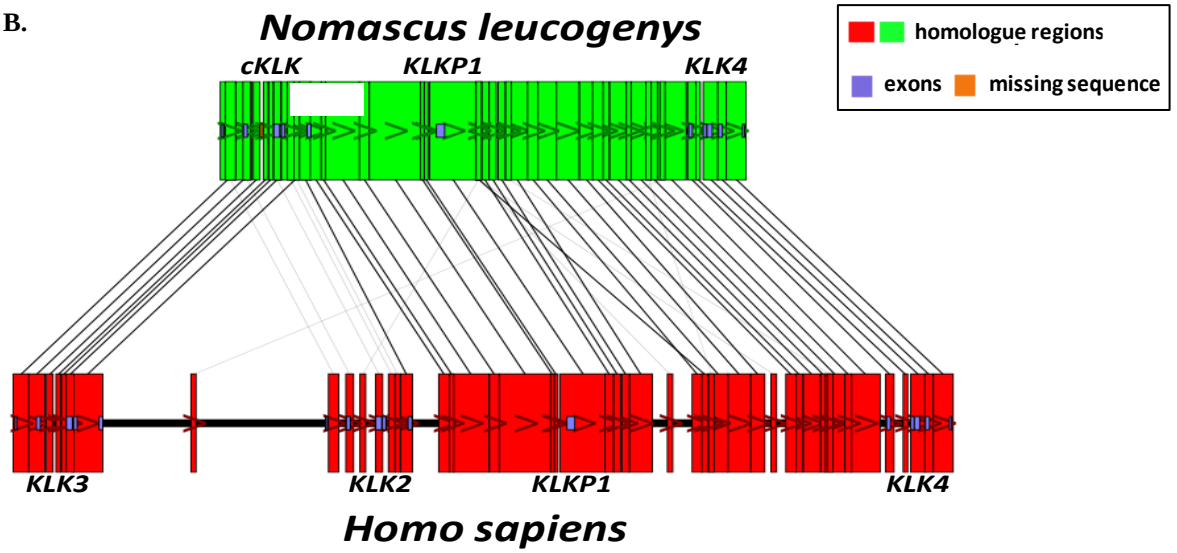
Species	Functional KLK number	Repeat Units[a]			Mating System[b]	Residual testis size[c]	Semen coagulation rating[d]
		SEMG1	SEMG2	Total			
Hsa	2	6	8	14	UM	-0.251	2
Ptr	2	11	5	16	MM	0.326	4
Ppa	2	9	8	17	MM	0.430	4
Ggo	1	5	6	11	UM	-0.622	2
Ppy	2	13	8	21	MM	-0.335	3
Nle.	1	6	8	14	UM	unknown	2
Hylobates sp.	1	3	7	10	UM	-0.471	2
Mmu	1	4	10	14	MM	0.320	3
Pan	1	11	4	15	MM	0.279	3
Cae	2	7	14	21	MM	0.238	
Cgu	0	8	10	18	UM	0.272	
Cja	0	5	6	11	UM	-0.366	2
Saguinus sp.	0	9	-	9	amb	-0.355	2
Saimiri sp.	1	10	-	10	MM	0.204	
Aotus sp.	0	-	6	6	UM	-1.066	2
Lca	1	-	7	7	MM	0.222	4
Callicebus sp.	0	7	6	13	UM	-0.114	

[a] Expected repeat units number of SEMGs protein (Jensen-Seaman and Li 2003; Hurle et al. 2007).
[b] Consensus mating system (Wlasiuk and Nachman 2010).
[c] Combined data from adult males (Anderson et al. 2004; Dixson and Anderson 2004; Wlasiuk and Nachman 2010).
[d] Semen coagulation is rated on a four-point scale (Dixson and Anderson 2002), with 1 reflecting no coagulation and 4 reflecting the production of a solid copulatory plug.
UM – Unimale or monoandrous
MM – Multimale or polyandrous
Amb - ambiguous

A.



B.



C.

Figure S1 - *KLK3-KLK2* gene fusion event in *Gorilla gorilla*, *Nomascus leucogenys* and *Hylobates sp*. Schematic representation of *G.* gorilla **(A)** or *N. leucogenys* **(B)** genomic sequence alignments with the *Homo sapiens* reference sequence (*KLK3* to *KLK4*). BlastN hits are represented as boxes joined with a line. Lighter lines indicate a non-optimal hit in one of the regions. Insertions and deletions cause a lack of correspondence between sequences. **(C)** Gene fusion event confirmed in *G. gorilla* and *Hylobates sp.* by PCR assay with gene- specific primers for *KLK3* (exon 4) and for *KLK2* (exon 5). The gene fusion product was confirmed in both species by sequencing of the resulting amplicons.

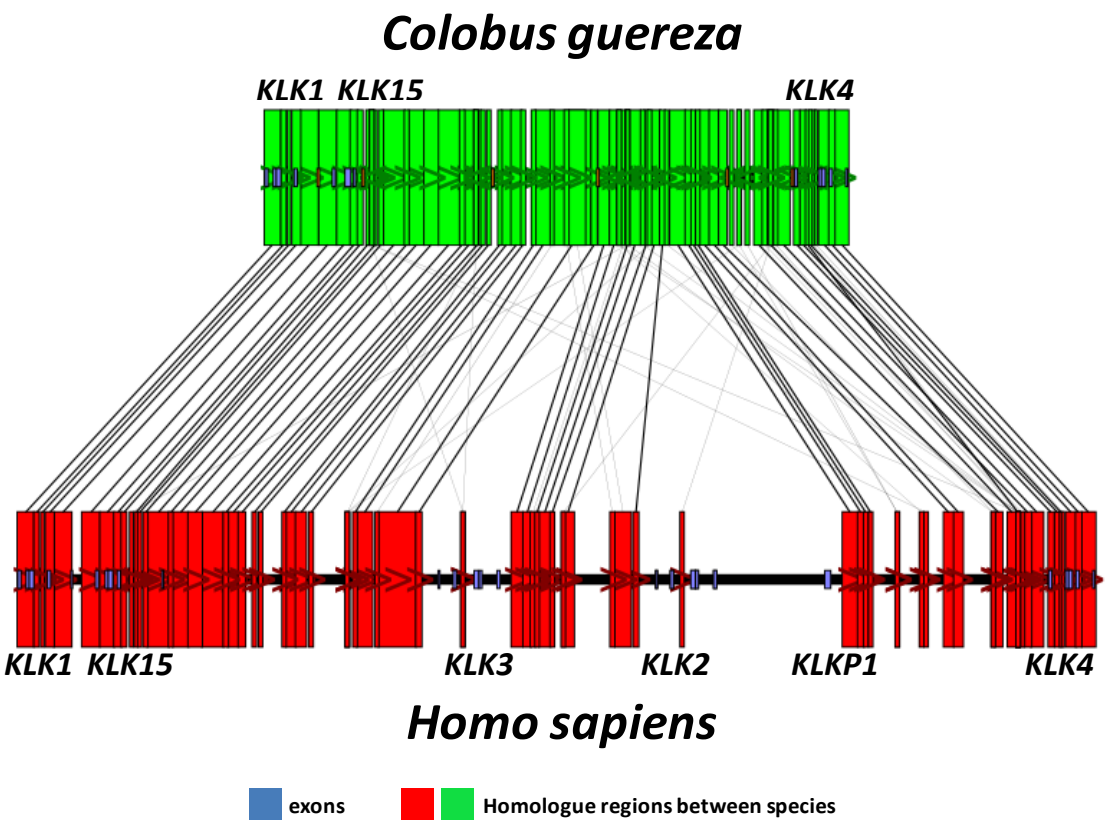


Figure S2 - Genomic sequence alignments of the orthologous genomic fragment spanning *KLK1*to *KLK4* in *Colobus guereza* and *Homo sapiens* (green and red, respectively). BlastN hits are represented as boxes joined with a line. Lighter lines indicate a non-optimal hit in one of the species. Insertions and deletions cause lack of correspondence between sequences.

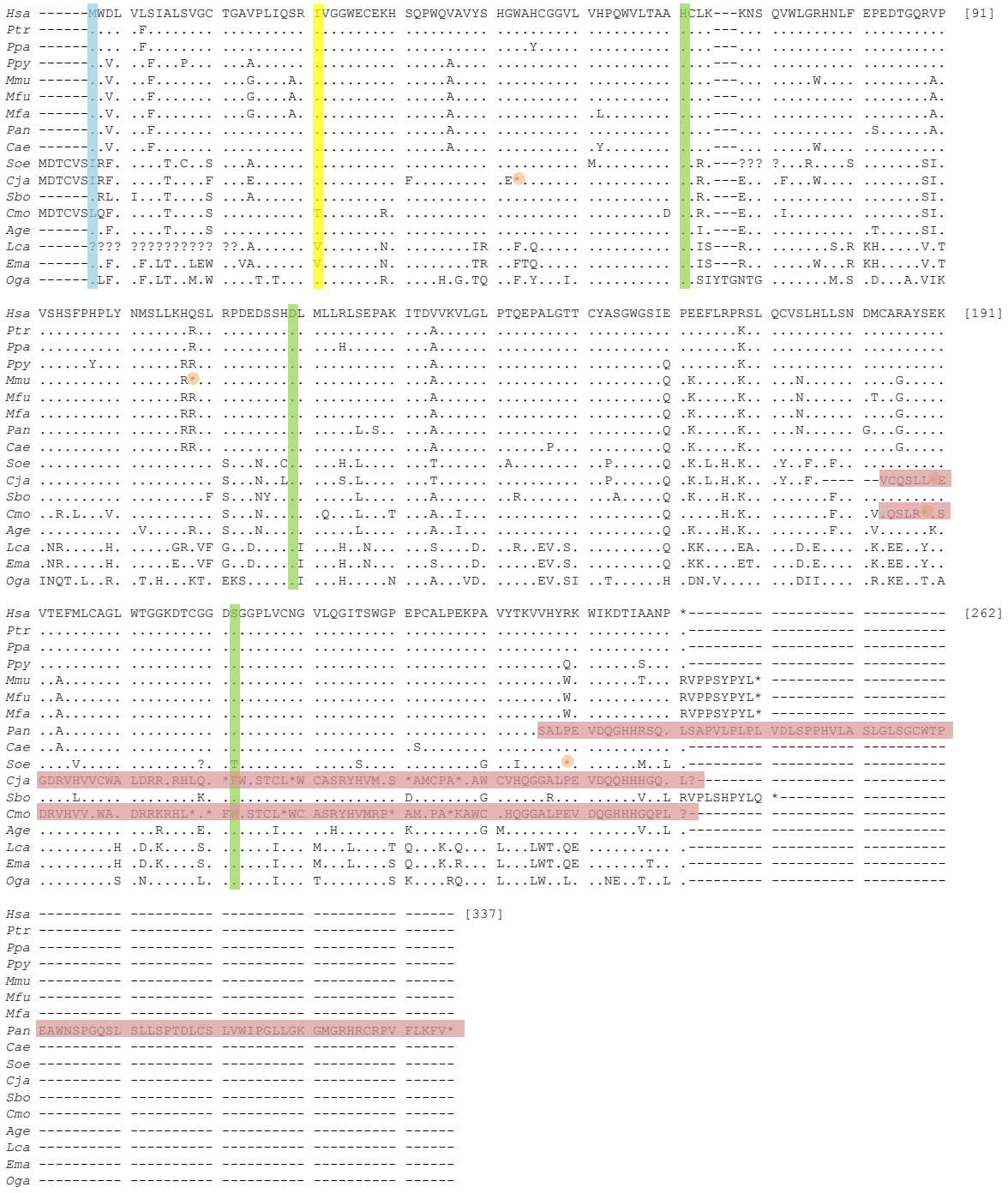


Figure S3 – KLK2 protein alignment identifying deleterious mutations. ( ) Start codon; ( ) Catalytic triad residues; ( ) Activation site; ( ) Frameshift; ( ) Premature STOP codons; (?) Missing data.

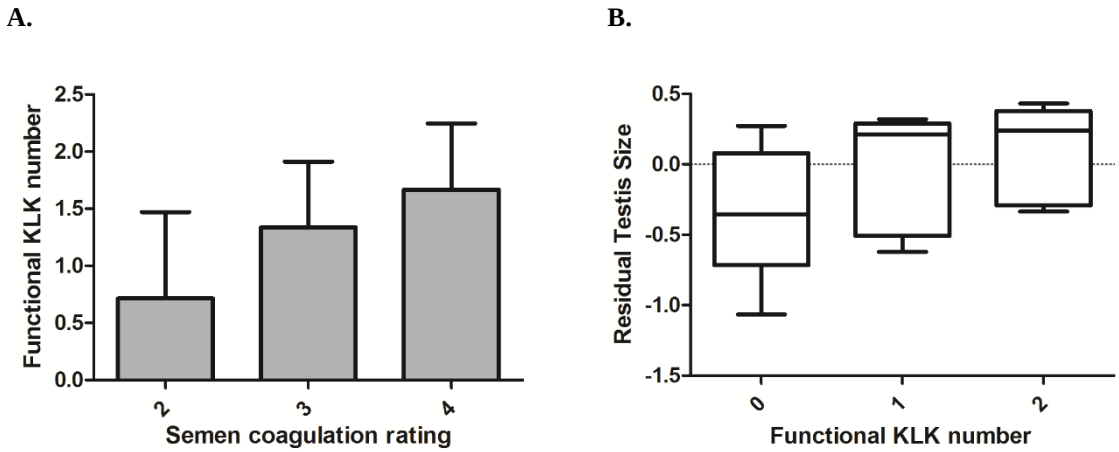


Figure S4 – Evolution of primate *KLK2* and *KLK3* related to reproductive traits. A) Correlation between the presence of functional KLK2 and KLK3 with semen coagulation rating. Semen coagulation is rated on a four-point scale (Dixson and Anderson 2002), with 1 reflecting no coagulation and 4 reflecting the production of a solid copulatory plug. B) Correlation of residual testis size (Anderson et al. 2004; Dixson and Anderson 2004; Wlasiuk and Nachman 2010) with the presence of functional KLK2 and KLK3.

References

Anderson MJ, Hessel JK, Dixson AF. 2004. Primate mating systems and the evolution of immune response. J Reprod Immunol 61:31-38.

Dixson AF, Anderson MJ. 2004. Sexual behavior, reproductive physiology and sperm competition in male mammals. Physiol Behav 83:361-371.

Dixson AL, Anderson MJ. 2002. Sexual selection, seminal coagulation and copulatory plug formation in primates. Folia Primatol (Basel) 73:63-69.

Hurle B, Swanson W, Program NCS, Green ED. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome Res 17:276-286.

Jensen-Seaman MI, Li WH. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. J Mol Evol 57:261-270.

Wlasiuk G, Nachman MW. 2010. Promiscuity and the rate of molecular evolution at primate immunity genes. Evolution 64:2204-2220.

Appendix B - Supplementary Material Paper II

Adaptive Evolution Favoring *KLK4* Downregulation in East Asians

Mol. Biol. Evol. 2015 *Epub ahead of print*

Supplementary Table S1 – Neutrality statistics data presented as –log10 of empirical ranked scores from 1000 Genomes Selection Browser (http://hsb.upf.edu/) for chr19:51353000-51461000 (GRCh37/hg19) region.

Due to the large size of the table this data is provided in digital format (Supplementary Tables – Paper II).

Supplementary Table S2 – Nonsynonymous and nonsense SNPs identified for the *KLK3-KLK5* segment in the 1000G data.

SNP ID	Protein	Frequency			Residue	SIFT	Polyphen
		ASN	CEU	YRI			
rs61752561	KLK3	0	0.035	0	D102N	Tolerated	Benign
rs61750343		0	0	0.006	R125C	Deleterious	Possibly damaging
rs182759459		0	0.006	0	E131K	Tolerated	Benign
rs2003783		0.094	0.106	0.136	L132I	Tolerated	Benign
rs17632542		0	0.112	0	I179T	Deleterious	Benign
rs61729813		0	0.035	0	S210W	Deleterious	Benign
rs149376489		0.003	0	0	R250Q	Tolerated	Benign
rs74478031	KLK2	0	0	0.011	Q22R	Deleterious	Possibly damaging
rs139063242		0	0.006	0	G52D	Deleterious	Probably damaging
rs181308768		0	0.006	0	S93R	Tolerated	Benign
rs140321127		0.003	0	0	K245N	Tolerated	Possibly damaging
rs198977		0.185	0.276	0.477	R250W	Tolerated	Benign
rs60268688		0	0	0.091	D255A	Deleterious	Benign
rs202006058	KLK4	0.003	0	0	G209E	Deleterious	Probably damaging
rs2569527		0	0.006	0	Q197H	Deleterious	Benign
rs34626614		0	0	0.074	G159D	Tolerated	Possibly damaging
rs104894704		0	0	0.011	W153*	-	-
rs182706497		0	0.012	0	R104Q	Tolerated	Benign
rs145560168		0	0.006	0	G80R	Tolerated	Possibly damaging
rs1654551		0.094	0.071	0.307	S22A	Tolerated	Benign
rs146723372	KLK5	0	0	0.006	R268Q	Tolerated	Benign
rs2232534		0	0.006	0	I172V	Tolerated	Benign
rs200140569		0.003	0	0	R164H	Tolerated	Benign
rs117067639		0	0.006	0	R117H	Tolerated	Probably damaging
rs201524467		0.003	0	0	P79L	Deleterious	Probably damaging
rs2232532		0	0.006	0	G55R	Tolerated	Benign

Supplementary Table S3 –Summary statistics of the *KLK3-KLK5* cluster segment from 1000G data.

Ancestry	Population	[a]N	*KLK3* (5849bp)						*KLK2* (7391bp)						*KLKP1* (14303bp)					
			[b]S	[c]θ_W	[d]π	[e]*D*	[f]*D**	[g]*H*	S	θ_W	π	*D*	*D**	*H*	S	θ_W	π	*D*	*D**	*H*
EAS	CHB	194	40	11.71	20.24	2.13	0.28	-0.13	36	8.42	8.08	-0.12	-0.28	-0.26	64	7.66	3.49	-1.65	-0.67	-2.75
	JPT	178	42	12.47	19.85	1.75	0.40	-0.43	35	8.31	8.02	-0.10	-0.65	-0.15	35	4.25	3.40	-0.58	-0.30	-1.29
	CHS	200	41	11.94	19.08	1.75	0.01	-0.90	34	7.91	7.95	0.01	0.65	-0.11	63	7.50	4.71	-1.12	0.86	-2.35
EUR	GBR	178	51	15.15	20.28	1.02	0.50	-0.68	39	9.26	10.61	0.42	-0.69	-0.13	68	8.26	7.23	-0.38	0.60	-1.33
	FIN	186	51	15.03	18.24	0.64	1.01	-1.23	34	8.01	10.85	1.03	1.02	0.08	62	7.47	7.19	-0.12	1.30	-1.45
	IBS	28	38	16.70	20.84	0.92	1.59	-0.56	26	9.13	10.72	0.63	-0.26	0.45	53	9.52	7.13	-0.95	-2.30	-1.35
	TSI	196	52	15.19	20.49	1.04	0.51	-0.57	37	8.64	11.72	1.04	-0.55	0.16	69	8.24	7.51	-0.27	-0.45	-1.22
AFR	ASW	122	69	21.94	20.21	-0.25	-0.74	-0.18	67	17.02	14.51	-0.47	-2.14	0.61	125	16.25	18.12	0.37	0.71	0.63
	LWK	194	70	20.48	19.18	-0.19	0.84	-0.31	70	16.37	14.52	-0.34	0.43	0.13	136	16.27	18.57	0.44	1.37	0.40
AMR	CLM	120	53	16.90	22.58	1.05	1.35	-0.35	51	13.00	12.10	-0.22	0.60	0.41	99	12.91	11.78	-0.28	-0.05	-0.13
	MXL	132	64	20.05	21.55	0.23	-1.03	-0.44	58	14.52	10.45	-0.88	-3.30	0.05	117	14.99	10.31	-1.01	-2.08	-0.40
	PUR	110	64	19.45	21.30	0.30	0.10	-.061	53	13.73	12.27	-0.34	-1.50	0.35	102	13.52	11.56	-0.47	-0.19	-0.32

[a]Number of chromosomes.

[b]Number of segregating sites.

[c]Watterson's estimator of θ (Watterson 1975) per base pair (x10-4).

[d]Nucleotide diversity per base pair (x10-4).

[e]Tajima's D statistic (Tajima 1989).

[f]Fu and Li D* (Fu 1997).

[g]Fay and Wu's H test (Fay and Wu 2000; Zeng et al. 2006).

Significant *P*-values < 0.05 according to the constant size model are underlined.

Supplementary Table S3 (Cont)

Ancestry	Population	N	*KLK4* (4387bp)						*KLK5* (9786bp)					
			S	θ_W	π	*D*	*D**	*H*	S	θ_W	π	*D*	*D**	*H*
EAS	CHB	194	25	9.75	10.72	0.27	0.57	0.02	48	8.40	10.18	0.63	0.35	-1.82
	JPT	178	25	9.90	10.38	0.14	-0.71	-0.02	48	8.52	9.91	0.49	0.38	-2.26
	CHS	200	24	9.32	12.44	0.92	0.50	0.36	51	8.87	9.28	0.14	0.46	-2.23
EUR	GBR	178	30	11.88	16.08	1.01	0.09	0.17	54	9.59	12.77	1.00	0.11	-2.10
	FIN	186	26	10.22	15.88	1.55	1.06	0.39	54	9.51	13.44	1.24	-0.16	-1.92
	IBS	28	21	12.30	15.94	1.05	0.17	-0.30	44	11.55	13.17	0.52	0.57	-2.00
	TSI	196	27	10.52	16.01	1.46	1.52	0.09	55	9.60	11.76	0.67	0.11	-2.41
AFR	ASW	122	35	14.84	17.78	0.60	0.80	-0.17	84	15.96	15.36	-0.12	-0.19	-0.68
	LWK	194	38	14.83	18.12	0.65	0.18	-0.30	93	16.27	15.62	-0.12	0.33	-0.86
AMR	CLM	120	31	13.18	17.35	0.95	0.28	0.34	60	11.44	13.75	0.64	0.47	-1.54
	MXL	132	32	13.37	16.39	0.67	-0.71	0.26	59	11.05	14.02	0.84	-1.09	-1.21
	PUR	110	30	12.97	15.75	0.65	0.25	0.09	61	11.82	13.49	0.45	-0.85	-1.47

Supplementary Table S4 – Sanger sequenced regions and HapMap Phase I/II samples.

Gene	Region (GRCh37/hg19)
KLK3	chr19: 51358238-51358859
	chr19: 51359064-51359988
	chr19: 51360882-51361876
	chr19: 51362745-51363368
KLK2	chr19: 51376708-51376823
	chr19: 51377493-51378534
	chr19: 51379488-51380823
	chr19: 51381126-51381811
KLKP1	chr19: 51385330-51385955
	chr19: 51391176-51392415
	chr19: 51397905-51399140
KLK4	chr19: 51409600-51409874
	chr19: 51410019-51410450
	chr19: 51411434-51413535
	chr19: 51413604-51414080
KLK5	chr19: 51446559-51447226
	chr19: 51451645-51452580
	chr19: 51451427-5145770
	chr19: 51453387-51456344

Population	Sample ID
CHB (15)	NA18532, NA18537, NA18545, NA18547, NA18562, NA18563, NA18572, NA18573, NA18576, NA18577, NA18579, NA18593, NA18603, NA18611, NA18623
JPT (15)	NA18940, NA18943, NA18944, NA18947, NA18949, NA18951, NA18952, NA18956, NA18959, NA18968, NA18974, NA18978, NA18994, NA19000, NA19012
CEU (10)	NA11830, NA11995, NA12006, NA12144, NA12249, NA12716, NA12751, NA12812, NA12814, NA12873
YRI (11)	NA18522, NA18853, NA18861, NA18871, NA19102, NA19138, NA19160, NA19171, NA19209, NA19093, NA19207

Supplementary Table S5 – Variants not detected by the 1000G project phase I.

Chromosome	Chromosome Position (GRCh37/hg19)	SNP ID	Gene	Population
19	51380601	rs8105211	*KLK2*	ASN; CEU; YRI
19	51380814	-	*KLK2*	ASN
19	51381276	-	*KLK2*	ASN
19	51391651	-	*KLKP1*	ASN
19	51409860	-	*KLK4*	CEU
19	51409977-51410044	esv1633192	*KLK4*	ASN; CEU; YRI
19	51412263	-	*KLK4*	YRI
19	51412321	-	*KLK4*	YRI
19	51412418	-	*KLK4*	ASN
19	51412882	-	*KLK4*	ASN
19	51413871	-	*KLK4*	YRI
19	51455406	-	*KLK5*	ASN
19	51456172	rs79633852	*KLK5*	ASN; CEU; YRI

Supplementary Table S6 – Summary statistics of the *KLK3-KLK5* cluster segment as estimated in our Sanger sequencing.

	ASN							CEU							YRI						
	[a]N	[b]S	[c]θ_W (SD)	[d]π (SD)	[e]*D*	[f]*D**	[g]*H*	N	S	θ_W (SD)	π (SD)	*D*	*D**	*H*	N	S	θ_W (SD)	π (SD)	*D*	*D**	*H*
KLK3 (3166bp)	60	19	21.84 (±4.43)	21.84 (±0.92)	<u>2.15*</u>	<u>1.68†*</u>	-0.37	20	20	17.81 (±7.05)	22.84 (±1.72)	1.07	<u>1.32*</u>	-0.38	22	24	20.80 (±7.91)	19.81 (±2.35)	-0.18	0.21	-0.36
KLK2 (3181bp	60	12	8.09 (±3.10)	7.46 (±1.50)	-0.22	0.35	<u>-2.30*</u>	20	14	12.41 (±5.20)	16.01 (±1.00)	1.09	0.78	0.01	22	22	18.97 (±7.30)	15.24 (±1.40)	-0.74	-0.94	0.25
KLKP1 (3102bp	60	7	4.84 (±2.18)	3.11 (±0.82)	-0.91	-1.19	<u>-5.36†*</u>	20	15	13.63 (±5.65)	19.00 (±3.00)	1.45	1.18	0.15	22	30	26.53 (±9.82)	28.75 (±1.94)	0.32	-0.46	0.72
KLK4 (3297bp)	60	19	12.36 (±4.26)	8.79 (±1.41)	-0.89	-1.57	-0.02	20	15	12.82 (±0.32)	14.53 (±1.28)	0.49	0.49	0.27	22	22	18.30 (±7.04)	18.62 (±2.12)	0.06	0.34	0.18
KLK5 (3663bp)	60	13	7.61 (±2.86)	9.16 (±0.71)	0.59	-0.09	-1.71	20	9	6.93 (±3.18)	8.87 (±1.74)	0.96	0.86	<u>-2.07*</u>	22	14	10.48 (±4.34)	12.76 (±0.95)	0.78	1.13	-0.37

[a]Number of chromosomes.
[b]Number of segregating sites.
[c]Watterson's estimator of θ (Watterson 1975) per base pair (x10-4).
[d]Nucleotide diversity per base pair (x10-4).
[e]Tajima's D statistic (Tajima 1989).
[f]Fu and Li D* (Fu 1997).
[g]Fay and Wu's H test (Fay and Wu 2000; Zeng et al. 2006).
Significant *P*-values < 0.05 according to the constant size model are underlined.
†Significant *P*-values < 0.05 according to Gravel model (Gravel et al. 2011) with recombination.
* Significant *P*-values < 0.05 according to Laval model (Laval et al. 2010) with recombination.

Supplementary Table S7 – iHS statistic for the 70 kb target region (chr19:51378273-51451045) in the ASN population.

Chromosome	Position (GRCh37/hg19)	SNP ID	Standardized |iHS|
19	51378725	rs1506685	0.760826
19	51378986	rs198971	0.320343
19	51379131	rs10403133	0.24359
19	51379556	rs11549920	1.35274
19	51379893	rs198972	1.32091
19	51380110	rs6070	0.993009
19	51380445	rs8105275	0.98984
19	51380602	rs198974	0.473825
19	51380815	rs198975	0.473825
19	51381083	rs8108845	1.02074
19	51381126	rs198976	0.473668
19	51381777	rs198977	1.16083
19	51383072	rs198978	0.604843
19	51385253	rs16987929	1.2164
19	51385402	rs8103659	1.21537
19	51388878	rs198958	0.700013
19	51389560	rs198957	0.495414
19	51390353	rs198956	0.497751
19	51390809	rs7256586	1.03071
19	51393118	rs1354774	1.15457
19	51393390	rs75380847	0.477157
19	51394286	rs2739482	0.629016
19	51396211	rs2739486	1.28641
19	51396580	rs998771	0.749558
19	51396922	rs61044983	1.31645
19	51397182	rs2659112	0.750004
19	51398071	rs8105985	1.38211
19	51398841	rs59618192	1.3406
19	51398888	rs1629856	0.880632
19	51400676	rs2739493	0.996974
19	51400812	rs2659107	1.55457
19	51400836	rs1560719	0.867248
19	51401807	rs3875143	1.50944
19	51402682	rs1654513	1.23174
19	51404513	rs2569536	2.18027
19	51405475	rs2739496	1.2985
19	51405982	rs806019	1.97009
19	51406129	rs11881354	1.47197
19	51406153	rs11881373	1.38777
19	51406201	rs7507565	1.58032
19	51406531	rs1090649	1.74605
19	51406618	rs2569531	2.17343
19	51407098	rs2739497	2.49244
19	51407293	rs56352322	2.49459
19	51407925	rs1701925	2.58113
19	51408758	rs1701926	1.6866
19	51408842	rs1701927	1.74944
19	51409763	rs1139132	2.15163
19	51409803	rs1654556	2.52879
19	51409876	rs12150961	2.10599
19	51410471	rs2235091	1.86394
19	51410772	rs1654554	1.68539
19	51411116	rs1654553	1.68134
19	51411329	rs1701929	2.83352
19	51411388	rs2979451	1.6505
19	51412122	rs117475014	0.124992

19	51412315	rs2242670	1.00889
19	51412326	rs2978643	2.0562
19	51412666	rs1654552	0.903704
19	51412668	rs1654551	1.06305
19	51413328	rs198968	1.71822
19	51413790	rs2242669	1.67238
19	51413802	rs198969	1.17784
19	51413906	rs2978642	1.69997
19	51414965	rs2979452	1.65573
19	51415150	rs2664152	1.54864
19	51415252	rs2664153	1.7979
19	51415515	rs118154507	1.08702
19	51418250	rs55700942	1.50025
19	51419062	rs2659077	0.918785
19	51419546	rs1701930	1.55512
19	51419669	rs1701931	1.55512
19	51419694	rs1701932	1.46185
19	51420025	rs1701933	1.46108
19	51420119	rs34225434	1.46108
19	51420319	rs78378001	1.00843
19	51420820	rs2659078	0.886996
19	51420996	rs62113140	1.46113
19	51421056	rs56311033	1.46113
19	51421096	rs10401284	1.42221
19	51421172	rs10425823	1.51356
19	51421255	rs55933733	1.61608
19	51421491	rs2659079	1.5121
19	51421833	rs62113142	1.5218
19	51421883	rs10419776	0.412897
19	51421979	rs10420003	0.987229
19	51422216	rs268923	1.49214
19	51422616	rs268922	1.25062
19	51422658	rs73598979	1.00856
19	51422691	rs268921	1.38611
19	51422694	rs75883262	1.01105
19	51422877	rs10427094	1.07727
19	51423231	rs10401844	1.7027
19	51423272	rs10403448	2.4497
19	51423360	rs10403688	2.62828
19	51423383	rs10402459	1.64671
19	51423391	rs10402465	1.64671
19	51423546	rs8100631	1.37185
19	51423628	rs8101572	1.28299
19	51424075	rs1532904	2.18817
19	51424078	rs1532903	1.41785
19	51424110	rs1532902	1.41785
19	51424126	rs6509501	1.41001
19	51424383	rs8104307	1.57298
19	51424425	rs268919	0.812089
19	51424448	rs8104644	1.38862
19	51424484	rs8104329	2.16171
19	51424607	rs138684768	0.0233301
19	51424651	rs117837287	1.13183
19	51424854	rs268917	1.43946
19	51424890	rs870361	2.23027
19	51425404	rs7254626	2.49066
19	51425614	rs7255201	1.68072
19	51426253	rs7258794	1.33302
19	51427332	rs116958492	0.866571
19	51427885	rs17727736	0.589067
19	51428729	rs112561158	0.861658

19	51428793	rs113141458	0.590272
19	51428914	rs113485158	0.0157743
19	51429512	rs17800825	0.553558
19	51429589	rs8111289	2.30832
19	51429596	rs8110335	1.84862
19	51429766	rs8111539	2.2872
19	51429883	rs8113547	1.77494
19	51429959	rs8100471	2.28275
19	51430277	rs11665937	2.25353
19	51430285	rs4802759	2.25353
19	51430853	rs12461743	0.409778
19	51431159	rs1865069	1.66187
19	51431447	rs7250053	1.66187
19	51431516	rs7250378	1.76855
19	51431836	rs7255268	1.08365
19	51431860	rs6509503	1.08365
19	51431895	rs6509504	1.08365
19	51431906	rs6509505	1.70436
19	51432054	rs6509506	1.70436
19	51432547	rs73600813	0.491551
19	51432717	rs113870369	0.0829882
19	51433003	rs2659081	0.466229
19	51433046	rs2739400	0.571263
19	51433048	rs2739401	0.571263
19	51433092	rs2739402	0.00324212
19	51433234	rs117145941	0.38612
19	51433803	rs8099967	1.3272
19	51433910	rs1654548	0.353653
19	51433915	rs1701905	0.220944
19	51434270	rs2472258	0.220755
19	51434353	rs2456586	1.81066
19	51434398	rs79966016	0.383806
19	51434627	rs1654546	0.190778
19	51434783	rs17800874	1.26907
19	51435101	rs115458416	0.253172
19	51435131	rs114406218	0.253172
19	51435261	rs111504285	0.321438
19	51435281	rs2569524	0.253172
19	51435299	rs16988270	0.987032
19	51435437	rs77202994	0.124606
19	51436255	rs1701910	0.309952
19	51436940	rs75024748	0.264681
19	51437537	rs149622538	0.264681
19	51437776	rs6509507	0.993766
19	51438178	rs12459790	0.26041
19	51439359	rs12460497	0.256661
19	51439564	rs9304706	0.256661
19	51439569	rs10164366	0.256661
19	51440217	rs10401225	1.40799
19	51440560	rs1701942	1.77283
19	51440564	rs1701943	2.38079
19	51440632	rs144230446	0.324326
19	51440658	rs8113756	1.44562
19	51440662	rs8113484	1.44562
19	51441046	rs150986447	0.231681
19	51441058	rs8102743	0.614044
19	51441071	rs146825647	0.55102
19	51441268	rs112062248	1.86888
19	51441759	rs11084040	2.09146
19	51441807	rs8104441	1.33381
19	51441915	rs73932685	0.612944

19	51442108	rs6509508	1.60735
19	51442397	rs77522061	0.653359
19	51442699	rs268914	0.666325
19	51443194	rs268913	0.676822
19	51444467	rs1812619	1.5598
19	51445723	rs62115181	1.53669
19	51446123	rs2739408	1.46314
19	51446246	rs2739409	0.875525
19	51446273	rs145620611	0.875525
19	51446327	rs2659090	1.70395
19	51446660	rs2659092	0.30144
19	51447065	rs1701949	1.74769
19	51447954	rs1897604	2.36264
19	51448182	rs12979210	1.69198
19	51448185	rs4802761	0.944289
19	51448685	rs12462803	0.826378
19	51448904	rs12463293	2.20184
19	51449566	rs55924070	1.55023
19	51449664	rs268909	1.61505
19	51449806	rs268908	1.00229
19	51449901	rs268907	1.75459
19	51449964	rs268906	1.50565
19	51449969	rs12459543	0.974763
19	51450534	rs10409028	1.03567
19	51450929	rs268905	1.22115
19	51451043	rs2411333	0.84265

Supplementary Table S8 – DIND statistic for the 70 kb target region (chr19:51378273-51451045) in the ASN population.

DAF - derived allele frequency; iπA - ancestral intra-allelic nucleotide diversity; iπD - derived intra-allelic nucleotide diversity.

Chromosome	Position (GRCh37/hg19)	SNP ID	DAF	iπA	iπD	iπA/iπD
19	51378273	rs62113074	0.024	52.55	12.78	4.11
19	51378725	rs1506685	0.129	52.33	46.60	1.12
19	51378782	rs139053005	0.005	51.88	6.00	8.65
19	51378821	rs2664158	0.035	52.46	25.82	2.03
19	51378883	rs189213221	0.005	51.37	159.00	0.32
19	51378884	rs2664159	0.027	52.28	27.07	1.93
19	51378986	rs198971	0.591	56.44	47.53	1.19
19	51379010	rs143718354	0.003	51.87	0.00	NA
19	51379041	rs113183318	0.040	52.56	25.64	2.05
19	51379042	rs111859906	0.046	51.55	53.22	0.97
19	51379131	rs10403133	0.403	46.65	57.60	0.81
19	51379255	rs138807230	0.003	51.27	0.00	NA
19	51379257	rs78379394	0.003	51.86	0.00	NA
19	51379556	rs11549920	0.073	50.35	55.58	0.91
19	51379806	rs199554707	0.003	51.79	0.00	NA
19	51379893	rs198972	0.288	33.91	73.01	0.46
19	51380110	rs6070	0.215	33.61	74.19	0.45
19	51380124	rs6071	0.013	52.14	11.40	4.57
19	51380364	rs34964828	0.011	51.92	24.50	2.12
19	51380445	rs8105275	0.212	34.90	70.33	0.50
19	51380602	rs198974	0.780	74.37	32.69	2.28
19	51380815	rs198975	0.780	74.37	32.69	2.28
19	51380903	rs190951085	0.003	51.85	0.00	NA
19	51381083	rs8108845	0.215	34.31	71.25	0.48
19	51381126	rs198976	0.780	74.37	32.69	2.28
19	51381392	rs145123696	0.003	51.88	0.00	NA
19	51381764	rs140321127	0.003	51.84	0.00	NA
19	51381777	rs198977	0.185	34.45	69.25	0.50
19	51382351	rs141318172	0.005	51.89	26.00	2.00
19	51382393	rs145052361	0.008	52.07	3.33	15.62
19	51382797	rs183546409	0.003	51.52	0.00	NA
19	51382882	rs143825608	0.003	51.85	0.00	NA
19	51382949	rs148175263	0.011	52.03	14.17	3.67
19	51383072	rs198978	0.806	73.73	32.30	2.28
19	51383149	rs193241860	0.003	51.70	0.00	NA
19	51383350	rs187930	0.005	50.72	119.00	0.43
19	51383826	rs150678087	0.003	51.86	0.00	NA
19	51384180	rs198979	0.995	119.00	50.72	2.35
19	51384460	rs2664143	0.995	119.00	50.72	2.35
19	51384500	rs185608	0.005	50.72	119.00	0.43
19	51384556	rs2659119	0.005	50.72	119.00	0.43
19	51384957	rs185255779	0.008	51.44	55.33	0.93
19	51385134	rs146499870	0.003	51.78	0.00	NA
19	51385253	rs16987929	0.191	33.50	71.67	0.47
19	51385259	rs146805696	0.005	51.71	73.00	0.71
19	51385402	rs8103659	0.191	33.50	71.67	0.47
19	51385483	rs198964	0.005	50.72	119.00	0.43
19	51385548	rs198963	0.005	50.72	119.00	0.43
19	51385617	rs138142886	0.016	51.99	36.07	1.44
19	51386494	rs139254001	0.003	51.67	0.00	NA
19	51387089	rs193277178	0.003	51.75	0.00	NA
19	51387248	rs116857887	0.008	51.93	9.33	5.56
19	51387721	rs146721840	0.011	51.79	47.17	1.10
19	51387899	rs142870280	0.003	51.84	0.00	NA

19	51387902	rs151037935	0.003	51.85	0.00	NA
19	51388135	rs189211714	0.005	51.85	10.00	5.19
19	51388158	rs181249001	0.003	51.86	0.00	NA
19	51388220	rs185885795	0.005	51.85	10.00	5.19
19	51388506	rs141834577	0.011	51.71	53.33	0.97
19	51388863	rs186168236	0.003	51.82	0.00	NA
19	51388878	rs198958	0.809	73.57	32.37	2.27
19	51389560	rs198957	0.833	70.43	32.81	2.15
19	51390171	rs111793246	0.005	50.72	119.00	0.43
19	51390353	rs198956	0.833	70.43	32.81	2.15
19	51390434	rs2659117	0.005	50.72	119.00	0.43
19	51390508	rs198955	0.005	50.72	119.00	0.43
19	51390636	rs67755001	0.005	50.72	119.00	0.43
19	51390809	rs7256586	0.161	34.34	66.38	0.52
19	51391121	rs66651444	0.005	50.72	119.00	0.43
19	51391243	rs144646444	0.003	51.87	0.00	NA
19	51391656	rs2664144	0.995	119.00	50.72	2.35
19	51391710	rs67008636	0.005	50.72	119.00	0.43
19	51391757	rs73054367	0.005	50.72	119.00	0.43
19	51392135	rs187986221	0.011	52.15	7.17	7.28
19	51392659	rs150327247	0.008	51.93	9.33	5.56
19	51392728	rs137966185	0.027	51.45	57.22	0.90
19	51392751	rs149477592	0.008	51.91	32.00	1.62
19	51393118	rs1354774	0.167	32.81	70.43	0.47
19	51393222	rs187508921	0.003	51.82	0.00	NA
19	51393252	rs191966599	0.003	51.87	0.00	NA
19	51393390	rs75380847	0.108	38.44	64.39	0.60
19	51393757	rs181507414	0.003	51.75	0.00	NA
19	51393831	rs10408670	0.005	50.72	119.00	0.43
19	51393870	rs118189204	0.030	52.51	4.62	11.37
19	51394286	rs2739482	0.836	69.33	33.37	2.08
19	51394690	rs148459857	0.003	51.88	0.00	NA
19	51394769	rs117402995	0.008	52.06	6.67	7.81
19	51394787	rs186064740	0.003	51.83	0.00	NA
19	51394808	rs2664146	0.005	50.72	119.00	0.43
19	51395038	rs2659115	0.005	50.72	119.00	0.43
19	51395122	rs2664147	0.995	119.00	50.72	2.35
19	51395128	rs2659114	0.005	50.72	119.00	0.43
19	51395183	rs2739483	0.995	119.00	50.72	2.35
19	51395560	rs2664148	0.005	50.72	119.00	0.43
19	51395790	rs79866735	0.005	51.48	44.00	1.17
19	51395819	rs2739484	0.995	119.00	50.72	2.35
19	51396201	rs145388082	0.003	51.87	0.00	NA
19	51396211	rs2739486	0.164	33.37	69.33	0.48
19	51396395	rs187725546	0.011	49.77	2.67	18.67
19	51396580	rs998771	0.836	69.33	33.37	2.08
19	51396922	rs61044983	0.159	34.88	65.12	0.54
19	51397007	rs2739487	0.005	50.72	119.00	0.43
19	51397101	rs2659113	0.005	50.72	119.00	0.43
19	51397182	rs2659112	0.836	69.33	33.37	2.08
19	51397212	rs56149619	0.005	50.72	119.00	0.43
19	51397651	rs74554810	0.022	51.47	52.96	0.97
19	51397897	rs147472492	0.008	52.00	12.67	4.11
19	51398071	rs8105985	0.161	34.87	65.60	0.53
19	51398191	rs1701939	0.995	119.00	50.72	2.35
19	51398377	rs1654517	0.005	50.72	119.00	0.43
19	51398569	rs55654965	0.005	50.72	119.00	0.43
19	51398841	rs59618192	0.159	34.88	65.12	0.54
19	51398864	rs117673213	0.003	51.86	0.00	NA
19	51398878	rs56325314	0.005	50.72	119.00	0.43
19	51398883	rs146687143	0.032	46.84	48.62	0.96

19	51398888	rs1629856	0.836	69.33	33.37	2.08
19	51399049	rs2739491	0.995	119.00	50.72	2.35
19	51399309	rs112035754	0.005	50.72	119.00	0.43
19	51399496	rs2659110	0.005	50.72	119.00	0.43
19	51399614	rs2659108	0.997	0.00	51.15	0.00
19	51400676	rs2739493	0.820	72.66	33.78	2.15
19	51400812	rs2659107	0.164	33.37	73.34	0.46
19	51400836	rs1560719	0.836	73.34	33.37	2.20
19	51400876	rs143903993	0.011	52.10	10.00	5.21
19	51401528	rs150896065	0.019	51.95	39.14	1.33
19	51401617	rs188115696	0.003	51.86	0.00	NA
19	51401797	rs8104538	0.022	50.63	69.11	0.73
19	51401807	rs3875143	0.161	32.69	72.57	0.45
19	51402682	rs1654513	0.817	78.79	30.56	2.58
19	51402706	rs183567117	0.003	51.87	0.00	NA
19	51403072	rs188295377	0.003	51.87	0.00	NA
19	51403390	rs149904243	0.005	51.94	5.00	10.39
19	51403423	rs806024	0.040	49.43	65.70	0.75
19	51403498	rs149022391	0.030	52.51	4.62	11.37
19	51403543	rs188588201	0.011	51.69	61.33	0.84
19	51403972	rs76720283	0.011	50.66	9.00	5.63
19	51403995	rs146992379	0.008	51.93	9.33	5.56
19	51404220	rs1701941	0.960	65.70	49.43	1.33
19	51404513	rs2569536	0.196	29.17	83.39	0.35
19	51404577	rs138731506	0.005	51.94	5.00	10.39
19	51404753	rs187212182	0.003	51.45	0.00	NA
19	51404955	rs192006428	0.011	52.13	6.17	8.45
19	51405084	rs806023	0.032	50.20	52.38	0.96
19	51405187	rs806022	0.960	65.70	49.43	1.33
19	51405412	rs191624202	0.003	51.84	0.00	NA
19	51405475	rs2739496	0.796	83.72	28.53	2.93
19	51405673	rs806021	0.032	50.20	52.38	0.96
19	51405808	rs806020	0.960	65.70	49.43	1.33
19	51405982	rs806019	0.202	28.46	83.60	0.34
19	51406129	rs11881354	0.161	32.69	72.57	0.45
19	51406153	rs11881373	0.140	35.87	72.20	0.50
19	51406201	rs7507565	0.132	49.96	60.54	0.83
19	51406213	rs186644524	0.024	50.48	35.00	1.44
19	51406241	rs192534387	0.005	51.94	5.00	10.39
19	51406286	rs184386975	0.003	51.67	0.00	NA
19	51406432	rs118100566	0.030	52.51	4.62	11.37
19	51406531	rs1090649	0.798	83.60	28.46	2.94
19	51406613	rs1090648	0.040	49.43	65.70	0.75
19	51406618	rs2569531	0.161	32.69	72.57	0.45
19	51406691	rs1090647	0.040	49.43	65.70	0.75
19	51407098	rs2739497	0.177	31.76	77.18	0.41
19	51407100	rs185221495	0.005	51.96	8.00	6.49
19	51407293	rs56352322	0.177	31.76	77.18	0.41
19	51407925	rs1701925	0.183	31.14	77.35	0.40
19	51408243	rs145819982	0.032	50.49	18.52	2.73
19	51408255	rs188706694	0.005	51.15	22.00	2.32
19	51408269	rs180724848	0.005	51.55	106.00	0.49
19	51408608	rs2569530	0.040	49.43	65.70	0.75
19	51408758	rs1701926	0.798	83.60	28.46	2.94
19	51408842	rs1701927	0.793	84.78	27.99	3.03
19	51409387	rs182271026	0.997	0.00	51.50	0.00
19	51409763	rs1139132	0.164	32.13	73.52	0.44
19	51409803	rs1654556	0.210	27.64	84.40	0.33
19	51409876	rs12150961	0.140	35.90	74.26	0.48
19	51410171	rs73042387	0.005	51.75	10.00	5.18
19	51410329	rs202006058	0.003	51.82	0.00	NA

19	51410398	rs198965	0.003	51.45	0.00	NA
19	51410471	rs2235091	0.161	32.32	73.91	0.44
19	51410772	rs1654554	0.806	84.50	29.97	2.82
19	51411116	rs1654553	0.105	48.37	64.78	0.75
19	51411263	rs7255024	0.008	51.04	113.33	0.45
19	51411329	rs1701929	0.258	27.04	88.35	0.31
19	51411356	rs75987180	0.040	51.62	39.94	1.29
19	51411388	rs2979451	0.156	33.87	73.88	0.46
19	51411392	rs117798052	0.011	51.90	39.33	1.32
19	51411404	rs145306141	0.003	51.78	0.00	NA
19	51411435	rs76655235	0.003	51.71	0.00	NA
19	51411675	rs149728389	0.003	51.86	0.00	NA
19	51412122	rs117475014	0.054	53.19	19.04	2.79
19	51412315	rs2242670	0.806	93.62	33.20	2.82
19	51412326	rs2978643	0.067	48.80	76.29	0.64
19	51412416	rs73042402	0.011	51.00	97.17	0.52
19	51412666	rs1654552	0.097	49.21	59.33	0.83
19	51412668	rs1654551	0.094	37.88	70.69	0.54
19	51412839	rs77569647	0.032	50.06	62.88	0.80
19	51413328	rs198968	0.739	93.26	26.24	3.55
19	51413395	rs141181534	0.003	51.86	0.00	NA
19	51413790	rs2242669	0.172	36.43	90.11	0.40
19	51413802	rs198969	0.124	48.23	64.47	0.75
19	51413906	rs2978642	0.220	34.90	87.18	0.40
19	51414522	rs116888134	0.048	43.34	44.09	0.98
19	51414737	rs75873880	0.003	51.86	0.00	NA
19	51414965	rs2979452	0.097	37.67	77.74	0.48
19	51415150	rs2664152	0.145	46.88	67.91	0.69
19	51415252	rs2664153	0.212	31.61	95.70	0.33
19	51415450	rs184421915	0.005	51.22	4.00	12.81
19	51415515	rs118154507	0.067	41.14	66.51	0.62
19	51415705	rs148701654	0.008	50.90	3.33	15.27
19	51415791	rs3760734	0.013	50.44	32.00	1.58
19	51416042	rs115034065	0.013	50.74	47.60	1.07
19	51416277	rs117319693	0.022	52.36	16.21	3.23
19	51417294	rs191358802	0.003	51.46	0.00	NA
19	51417736	rs10424317	0.992	100.67	50.84	1.98
19	51417852	rs190593989	0.003	51.78	0.00	NA
19	51417958	rs7246794	0.003	51.27	0.00	NA
19	51418250	rs55700942	0.094	43.49	73.66	0.59
19	51418978	rs188524307	0.003	51.86	0.00	NA
19	51419016	rs150522181	0.003	51.87	0.00	NA
19	51419062	rs2659077	0.906	73.66	43.49	1.69
19	51419546	rs1701930	0.097	43.49	73.22	0.59
19	51419669	rs1701931	0.097	43.49	73.22	0.59
19	51419694	rs1701932	0.817	98.54	25.82	3.82
19	51419812	rs184968554	0.005	51.82	46.00	1.13
19	51420025	rs1701933	0.097	43.49	73.22	0.59
19	51420119	rs34225434	0.097	43.49	73.22	0.59
19	51420319	rs78378001	0.086	37.59	62.12	0.61
19	51420820	rs2659078	0.903	73.22	43.49	1.68
19	51420996	rs62113140	0.097	43.49	73.22	0.59
19	51421056	rs56311033	0.097	43.49	73.22	0.59
19	51421096	rs10401284	0.823	98.29	26.24	3.75
19	51421172	rs10425823	0.817	98.54	25.82	3.82
19	51421255	rs55933733	0.782	98.68	25.91	3.81
19	51421316	rs186188017	0.008	51.51	57.33	0.90
19	51421491	rs2659079	0.097	43.49	73.22	0.59
19	51421833	rs62113142	0.817	98.54	25.82	3.82
19	51421883	rs10419776	0.914	62.12	37.59	1.65
19	51421979	rs10420003	0.086	37.59	62.12	0.61

19	51422216	rs268923	0.097	43.15	74.41	0.58
19	51422616	rs268922	0.075	43.72	55.92	0.78
19	51422658	rs73598979	0.089	37.22	62.63	0.59
19	51422691	rs268921	0.836	93.95	26.06	3.60
19	51422694	rs75883262	0.089	37.22	62.63	0.59
19	51422877	rs10427094	0.091	36.90	65.14	0.57
19	51422962	rs116947194	0.005	51.18	7.00	7.31
19	51423231	rs10401844	0.785	95.13	24.79	3.84
19	51423272	rs10403448	0.215	24.79	95.13	0.26
19	51423360	rs10403688	0.226	24.42	95.51	0.26
19	51423383	rs10402459	0.790	95.06	24.85	3.83
19	51423391	rs10402465	0.790	95.06	24.85	3.83
19	51423546	rs8100631	0.831	94.06	25.39	3.70
19	51423603	rs150119603	0.011	52.13	6.17	8.45
19	51423604	rs117343646	0.008	52.05	6.67	7.81
19	51423628	rs8101572	0.078	43.56	55.86	0.78
19	51423641	rs183548434	0.003	51.86	0.00	NA
19	51423744	rs117098406	0.024	51.89	36.39	1.43
19	51424075	rs1532904	0.172	25.36	94.53	0.27
19	51424078	rs1532903	0.828	94.53	25.36	3.73
19	51424110	rs1532902	0.828	94.53	25.36	3.73
19	51424126	rs6509501	0.081	43.58	58.20	0.75
19	51424383	rs8104307	0.828	93.62	30.10	3.11
19	51424425	rs268919	0.946	50.57	47.06	1.07
19	51424448	rs8104644	0.844	94.52	30.31	3.12
19	51424484	rs8104329	0.156	30.08	93.44	0.32
19	51424607	rs138684768	0.051	42.44	34.91	1.22
19	51424619	rs141274704	0.027	48.86	29.09	1.68
19	51424651	rs117837287	0.091	36.90	65.14	0.57
19	51424854	rs268917	0.054	47.06	50.57	0.93
19	51424890	rs870361	0.145	30.17	91.16	0.33
19	51425251	rs117702669	0.003	51.88	0.00	NA
19	51425404	rs7254626	0.175	24.85	94.70	0.26
19	51425614	rs7255201	0.825	94.70	24.85	3.81
19	51425748	rs187040616	0.003	51.84	0.00	NA
19	51426253	rs7258794	0.054	46.92	48.37	0.97
19	51426271	rs75111430	0.970	44.98	48.68	0.92
19	51426285	rs114624700	0.030	48.68	44.98	1.08
19	51426779	rs182320116	0.005	51.92	14.00	3.71
19	51426813	rs187025895	0.003	51.69	0.00	NA
19	51427076	rs192361415	0.003	51.86	0.00	NA
19	51427332	rs116958492	0.051	47.08	69.25	0.68
19	51427571	rs144574553	0.003	51.71	0.00	NA
19	51427572	rs148021763	0.016	52.03	17.33	3.00
19	51427885	rs17727736	0.070	38.97	40.59	0.96
19	51428729	rs112561158	0.051	47.08	69.25	0.68
19	51428793	rs113141458	0.070	38.97	40.59	0.96
19	51428914	rs113485158	0.930	40.59	38.97	1.04
19	51429232	rs150060784	0.005	51.94	9.00	5.77
19	51429332	rs79735327	0.030	52.68	15.60	3.38
19	51429512	rs17800825	0.070	38.97	40.59	0.96
19	51429589	rs8111289	0.175	24.85	94.70	0.26
19	51429596	rs8110335	0.785	95.25	24.28	3.92
19	51429608	rs144627964	0.003	51.63	0.00	NA
19	51429703	rs181276920	0.003	51.75	0.00	NA
19	51429766	rs8111539	0.175	24.82	94.97	0.26
19	51429883	rs8113547	0.815	96.10	24.99	3.85
19	51429959	rs8100471	0.175	24.82	94.97	0.26
19	51430006	rs139175576	0.005	50.79	2.00	25.39
19	51430277	rs11665937	0.177	24.87	95.31	0.26
19	51430285	rs4802759	0.177	24.87	95.31	0.26

19	51430436	rs17714545	0.024	49.78	52.50	0.95
19	51430574	rs186172502	0.005	51.83	6.00	8.64
19	51430853	rs12461743	0.070	38.97	40.59	0.96
19	51431159	rs1865069	0.102	41.76	65.45	0.64
19	51431447	rs7250053	0.102	41.76	65.45	0.64
19	51431516	rs7250378	0.782	95.91	23.77	4.03
19	51431836	rs7255268	0.898	65.45	41.76	1.57
19	51431860	rs6509503	0.898	65.45	41.76	1.57
19	51431895	rs6509504	0.898	65.45	41.76	1.57
19	51431906	rs6509505	0.102	41.76	65.45	0.64
19	51432054	rs6509506	0.102	41.76	65.45	0.64
19	51432210	rs143614792	0.995	61.00	51.78	1.18
19	51432547	rs73600813	0.070	38.97	40.59	0.96
19	51432717	rs113870369	0.930	40.59	38.97	1.04
19	51433002	rs150417602	0.022	52.37	18.64	2.81
19	51433003	rs2659081	0.070	38.97	40.59	0.96
19	51433046	rs2739400	0.073	39.03	47.64	0.82
19	51433048	rs2739401	0.073	39.03	47.64	0.82
19	51433092	rs2739402	0.927	47.64	39.03	1.22
19	51433234	rs117145941	0.067	39.22	35.51	1.10
19	51433803	rs8099967	0.825	95.28	26.13	3.65
19	51433910	rs1654548	0.067	39.22	35.51	1.10
19	51433915	rs1701905	0.933	35.51	39.22	0.91
19	51434243	rs10409216	0.003	51.77	0.00	NA
19	51434270	rs2472258	0.933	35.51	39.22	0.91
19	51434353	rs2456586	0.113	42.37	70.45	0.60
19	51434398	rs79966016	0.067	39.22	35.51	1.10
19	51434627	rs1654546	0.933	35.51	39.22	0.91
19	51434783	rs17800874	0.801	96.58	29.02	3.33
19	51435101	rs115458416	0.930	41.32	39.09	1.06
19	51435131	rs114406218	0.930	41.32	39.09	1.06
19	51435261	rs111504285	0.070	39.09	41.32	0.95
19	51435281	rs2569524	0.930	41.32	39.09	1.06
19	51435299	rs16988270	0.836	94.79	33.94	2.79
19	51435437	rs77202994	0.067	53.16	27.21	1.95
19	51435606	rs67002911	0.032	48.51	37.80	1.28
19	51435724	rs188570150	0.008	51.97	20.67	2.51
19	51435736	rs1701906	0.008	51.67	61.33	0.84
19	51435779	rs184266168	0.003	51.88	0.00	NA
19	51435803	rs141297744	0.011	52.06	16.33	3.19
19	51435978	rs35047583	0.032	48.51	37.80	1.28
19	51436255	rs1701910	0.930	41.32	39.09	1.06
19	51436940	rs75024748	0.070	39.09	41.32	0.95
19	51437537	rs149622538	0.070	39.09	41.32	0.95
19	51437776	rs6509507	0.844	95.48	28.88	3.31
19	51438015	rs142560395	0.003	51.88	0.00	NA
19	51438178	rs12459790	0.070	39.09	41.32	0.95
19	51439205	rs192588396	0.003	51.75	0.00	NA
19	51439359	rs12460497	0.070	39.09	41.32	0.95
19	51439564	rs9304706	0.070	39.09	41.32	0.95
19	51439569	rs10164366	0.070	39.09	41.32	0.95
19	51440134	rs192681104	0.003	51.88	0.00	NA
19	51440217	rs10401225	0.086	44.40	61.31	0.72
19	51440237	rs140484129	0.005	50.78	2.00	25.39
19	51440264	rs145628364	0.003	51.68	0.00	NA
19	51440343	rs138496452	0.005	51.80	55.00	0.94
19	51440420	rs150678688	0.003	51.28	0.00	NA
19	51440560	rs1701942	0.124	49.34	54.92	0.90
19	51440564	rs1701943	0.177	44.10	69.37	0.64
19	51440632	rs144230446	0.070	39.09	41.32	0.95
19	51440658	rs8113756	0.086	44.40	61.31	0.72

19	51440662	rs8113484	0.086	44.40	61.31	0.72
19	51440753	rs1701945	0.048	48.14	48.49	0.99
19	51441046	rs150986447	0.500	70.15	27.30	2.57
19	51441058	rs8102743	0.484	66.39	31.78	2.09
19	51441071	rs146825647	0.070	39.52	46.64	0.85
19	51441268	rs112062248	0.879	69.28	44.08	1.57
19	51441682	rs190414825	0.003	51.86	0.00	NA
19	51441759	rs11084040	0.151	29.69	95.10	0.31
19	51441807	rs8104441	0.849	95.10	29.69	3.20
19	51441915	rs73932685	0.070	39.52	46.64	0.85
19	51441934	rs191258507	0.003	51.82	0.00	NA
19	51442108	rs6509508	0.852	94.69	29.70	3.19
19	51442397	rs77522061	0.070	39.52	46.64	0.85
19	51442534	rs148685704	0.013	51.34	84.20	0.61
19	51442699	rs268914	0.070	39.52	46.64	0.85
19	51443009	rs140916125	0.005	51.92	17.00	3.05
19	51443194	rs268913	0.070	39.52	46.64	0.85
19	51443488	rs80161131	0.003	51.51	0.00	NA
19	51443922	rs117578550	0.011	51.10	24.17	2.11
19	51443937	rs147098568	0.003	51.87	0.00	NA
19	51444189	rs192759906	0.003	51.48	0.00	NA
19	51444233	rs268911	0.970	33.05	48.58	0.68
19	51444239	rs268910	0.970	33.05	48.58	0.68
19	51444313	rs188302848	0.005	51.51	104.00	0.50
19	51444467	rs1812619	0.108	36.59	89.01	0.41
19	51444709	rs1812927	0.030	48.58	33.05	1.47
19	51444761	rs965601	0.003	51.87	0.00	NA
19	51444835	rs972921	0.030	48.58	33.05	1.47
19	51444987	rs972920	0.030	48.58	33.05	1.47
19	51445074	rs181668987	0.003	51.80	0.00	NA
19	51445178	rs186869888	0.008	51.74	52.67	0.98
19	51445424	rs36120506	0.032	48.51	37.80	1.28
19	51445506	rs35418865	0.032	48.51	37.80	1.28
19	51445543	rs146243744	0.011	52.14	7.33	7.11
19	51445723	rs62115181	0.860	95.40	31.82	3.00
19	51446123	rs2739408	0.911	54.53	50.12	1.09
19	51446228	rs2569523	0.965	42.49	49.24	0.86
19	51446246	rs2739409	0.070	39.84	50.90	0.78
19	51446273	rs145620611	0.070	39.84	50.90	0.78
19	51446327	rs2659090	0.105	35.68	89.17	0.40
19	51446530	rs2659091	0.965	42.49	49.24	0.86
19	51446660	rs2659092	0.930	50.90	39.84	1.28
19	51447065	rs1701949	0.075	50.08	58.64	0.85
19	51447270	rs141858477	0.005	51.97	3.00	17.32
19	51447353	rs2253655	0.965	42.49	49.24	0.86
19	51447954	rs1897604	0.142	31.78	95.63	0.33
19	51448144	rs139835369	0.005	51.95	13.00	4.00
19	51448182	rs12979210	0.858	95.63	31.78	3.01
19	51448185	rs4802761	0.070	39.84	50.90	0.78
19	51448685	rs12462803	0.067	40.31	50.99	0.79
19	51448904	rs12463293	0.202	31.38	95.23	0.33
19	51449038	rs142427219	0.003	51.86	0.00	NA
19	51449566	rs55924070	0.831	97.37	30.56	3.19
19	51449664	rs268909	0.105	36.64	90.96	0.40
19	51449806	rs268908	0.895	90.96	36.64	2.48
19	51449901	rs268907	0.113	36.11	93.62	0.39
19	51449916	rs12979237	0.027	51.05	58.51	0.87
19	51449964	rs268906	0.089	39.58	97.17	0.41
19	51449969	rs12459543	0.073	39.95	87.46	0.46
19	51450081	rs140217681	0.997	0.00	51.86	0.00
19	51450150	rs181274545	0.003	51.50	0.00	NA

19	51450337	rs11666803	0.035	50.21	81.67	0.61
19	51450351	rs62115183	0.038	50.28	78.62	0.64
19	51450491	rs186064448	0.003	51.86	0.00	NA
19	51450534	rs10409028	0.159	51.11	53.43	0.96
19	51450661	rs73045911	0.038	51.85	45.30	1.14
19	51450678	rs10409107	0.957	38.99	52.24	0.75
19	51450694	rs148233985	0.003	51.82	0.00	NA
19	51450835	rs141477292	0.003	51.87	0.00	NA
19	51450929	rs268905	0.121	52.00	47.72	1.09
19	51451043	rs2411333	0.194	51.82	49.92	1.04

Supplementary Table S9 – Candidate variants

Due to the large size of the table this data is provided in digital format (Supplementary Tables – Paper II).

Supplementary Table S10 – CNV and rs1654556 genotypes for the ASN HapMap Phase I/II samples included in our Sanger sequencing study dataset.

Sample ID	Genotype	
	CNV	rs1654556
NA18532	Ins/Del105	G/A
NA18537	Ins/Ins	G/G
NA18545	Ins/Del67	G/A
NA18547	Ins/Ins67	G/A
NA18562	Ins/Ins	G/G
NA18563	Ins/Ins	G/G
NA18572	Ins/Ins	G/G
NA18573	Ins/Ins	G/G
NA18576	Ins/Ins	G/G
NA18577	Ins/Ins	G/G
NA18579	Ins/Del67	G/A
NA18593	Ins/Ins	G/G
NA18603	Ins/Ins67	G/A
NA18611	Ins/Ins	G/G
NA18623	Ins/Ins	G/G
NA18940	Ins/Ins	G/G
NA18943	Ins/Ins	G/G
NA18944	Ins/Ins	G/G
NA18947	Ins/Ins	G/G
NA18949	Ins/Ins67	G/A
NA18951	Ins/Ins	G/G
NA18952	Ins67/Ins67	A/A
NA18956	Ins/Ins	G/G
NA18959	Ins/Ins67	G/A
NA18968	Ins/Ins	G/G
NA18974	Ins/Ins	G/G
NA18978	Ins/Ins	G/G
NA18994	Ins/Ins	G/G
NA19000	Ins/Ins67	G/A
NA19012	Ins/Ins	G/G

Supplementary Table S11 – Command lines used in the ms software.

Gravel model (Gravel et al. 2011) with recombination for ASN population.
ms N 100000 -s SEGSITES -r RECOMB_RATE BP -c 2 500 -G 142.30 -eN 0.0315 0.2546 -eN 0.06977 1.98 -eN 0.2025 1
Gravel model (Gravel et al. 2011) with recombination for CEU population.
ms N 100000 -s SEGSITES -r RECOMB_RATE BP -c 2 500 -G 112.674 -eN 0.0315 0.2546 -eN 0.06977 1.98 -eN 0.2025 1
Laval model (Laval et al. 2010) with recombination for ASN population
ms N 100000 -s SEGSITES -r RECOMB_RATE BP -c 2 500 -eG 0 39.7222 -eN 0.0414 0.9515
Laval model (Laval et al. 2010) with recombination for CEU population
ms N 100000 -s SEGSITES -r RECOMB_RATE BP -c 2 500 -eG 0 126.884 -eN 0.019 0.442

N = number of chromosomes

SEGSITES = Segregating sites observed for each population and for each gene independently.

RECOMB_RATE = recombination rate parameter as inferred from HapMap phase II data (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/latest/old_data/rates/) (McVean et al. 2004).

BP = Total base pairs (bp) anayzed per gene.

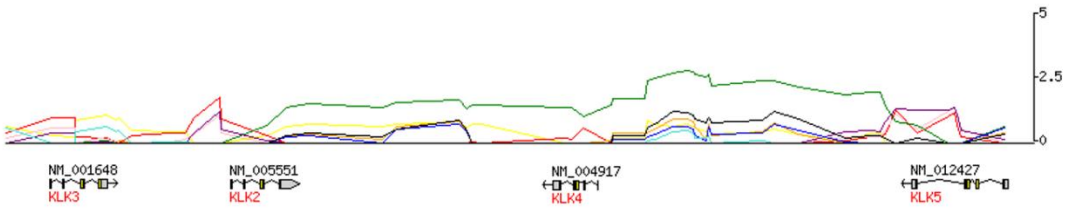
Supplementary Table S12 – Primers used for generation of luciferase reporter constructs for rs198968, rs17800874 and rs1654556.

	Forward primer	Reverse primer
rs198968	AATCTGTGCCTGCTTCCTGG	CCCGAGCAGAACAATTCGTT
rs17800874	AGTCTTTCTGTTGAGGTGGT	AGCATCTAATTGTTGGCTAC
rs1654556	TCTGGAATGGGACTTCCAAC	CCATGGAGGGAAAGCCATTT

Supplementary Table S13 – Primers used for cDNA amplification of *KLK3*, *KLK2*, *KLKP1*, *KLK4* and *KLK5* transcripts.

	Forward primer	Reverse primer
KLK3	GAATCGATTCCTCAGGCCAG	CAATAGGGGGTTGATAGGGG
KLK2	GCATCAAAGCCTTAGACCAG	ATGCCAGAACGTGAGGTGGAC
KLKP1	GCTCTCAGAGCAAAGTCTCC	GGGTGATGCAGTGAGCAGTA
	CTGGGGCATCTTAGAGCATC	ACCCAGGATGTGAAAGTTGC
KLK4	AACGAATTGTTCTGCTCGGG	CACTGCGAAGCAATGCTGAT
KLK5	AAAGTGCTTGGTGTCTGGCT	TCAACATCTCTGGGAAGGAATG

A



B

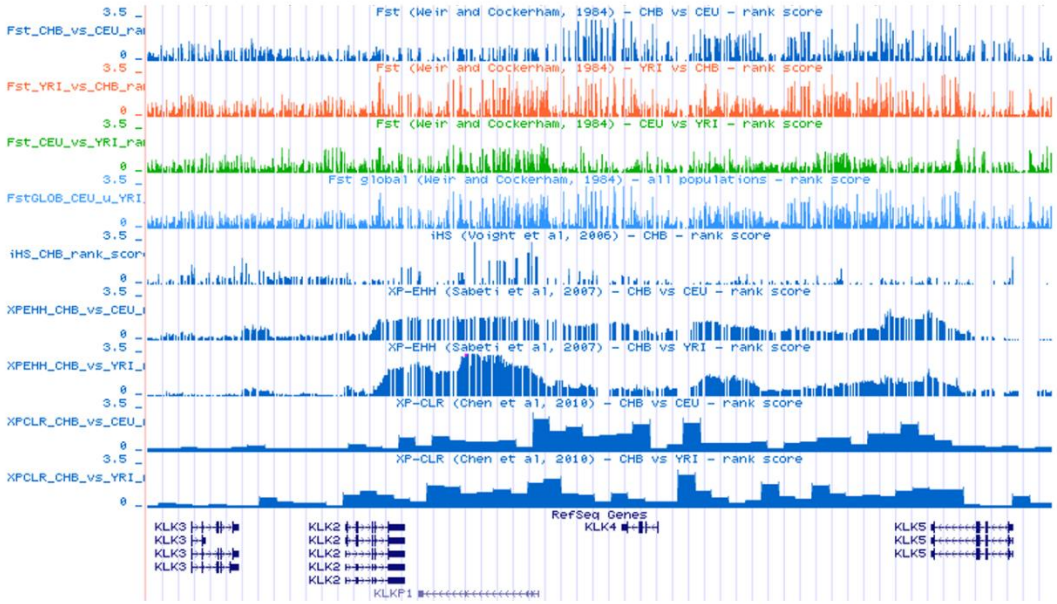


Figure S1 – Selection statistics for *KLK3-KLK5* locus. (A) Cross-Population Extended Haplotype Homozygosity (XP-EHH) plot from HGDP data for different continental populations as indicated by different color lines (http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/). East Asia is represented in green, South Asia in black, Europe in orange, Mideast in blue, Oceania in turquoise, America in yellow, Bantu in red and non-Bantu African populations in pink and purple. **(B)** 1000 Genomes Selection Browser view. Statistic tracks for pairwise F_{ST} for CHB *vs.* CEU, YRI *vs.* CHB and CEU *vs.* YRI, F_{ST} Global (CHB, CEU and YRI), integrated haplotype score (iHS) for CHB, cross-population extended haplotype homozygosity (XP-EHH) for CHB *vs.* CEU and YRI *vs.* CHB, and cross-population composite likelihood ratio (XP-CLR) for CHB *vs.* CEU and YRI *vs.* CHB. The statistics are presented as –log10 of empirical ranked scores (http://hsb.upf.edu/).

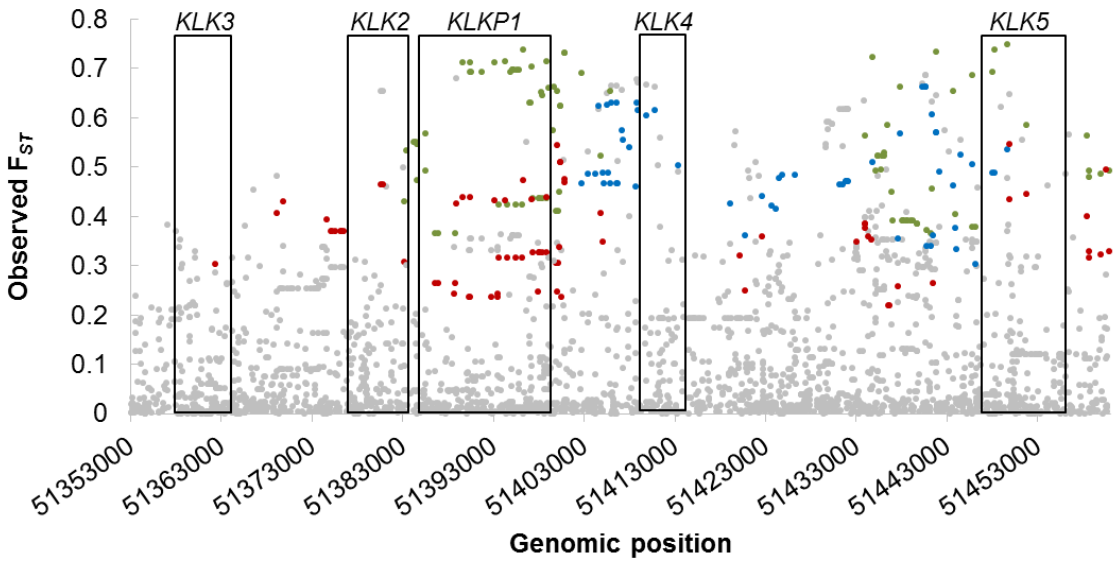


Figure S2 – Genetic population differentiation (F_{ST}) analysis for *KLK3-KLK5* locus of ASN *vs.* CEU, ASN *vs.* YRI and CEU *vs.* YRI populations. Genes' location is delimited by open boxes. SNPs with significant F_{ST} *P*-values (upper $P < 0.05$) are displayed in blue, green and red for ASN *vs.* CEU, ASN *vs.* YRI and CEU *vs.* YRI comparisons, respectively.

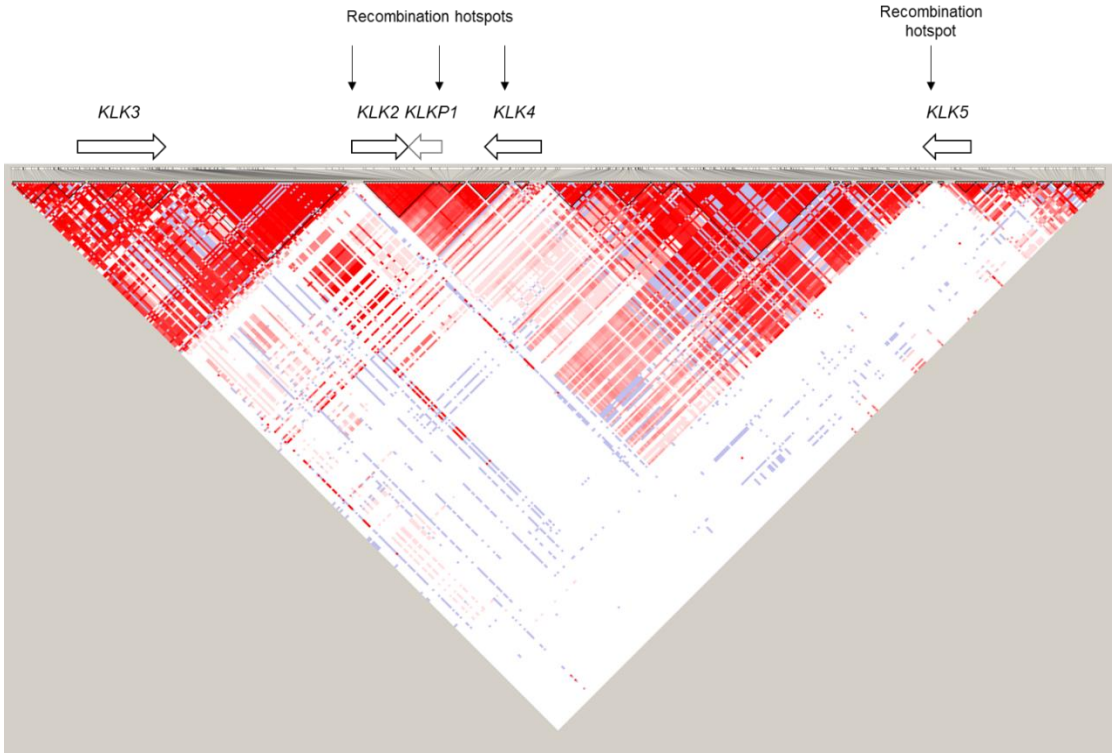


Figure S3 – Linkage disequilibrium plot of 1000G phase I data for *KLK3-KLK5* region in Asians. The image was generated using *Haploview* 4.2 software. The triangular units represent haplotype blocks as defined by Gabriel et al. 2002. The degree of LD between pair of markers is indicated by the |D'| statistic (|D'| = 1, bright red; |D'| < 1, shades of red). The relative positions of *KLK* genes are depicted by open arrows, and the relative positions of the recombination hotspots are also shown.

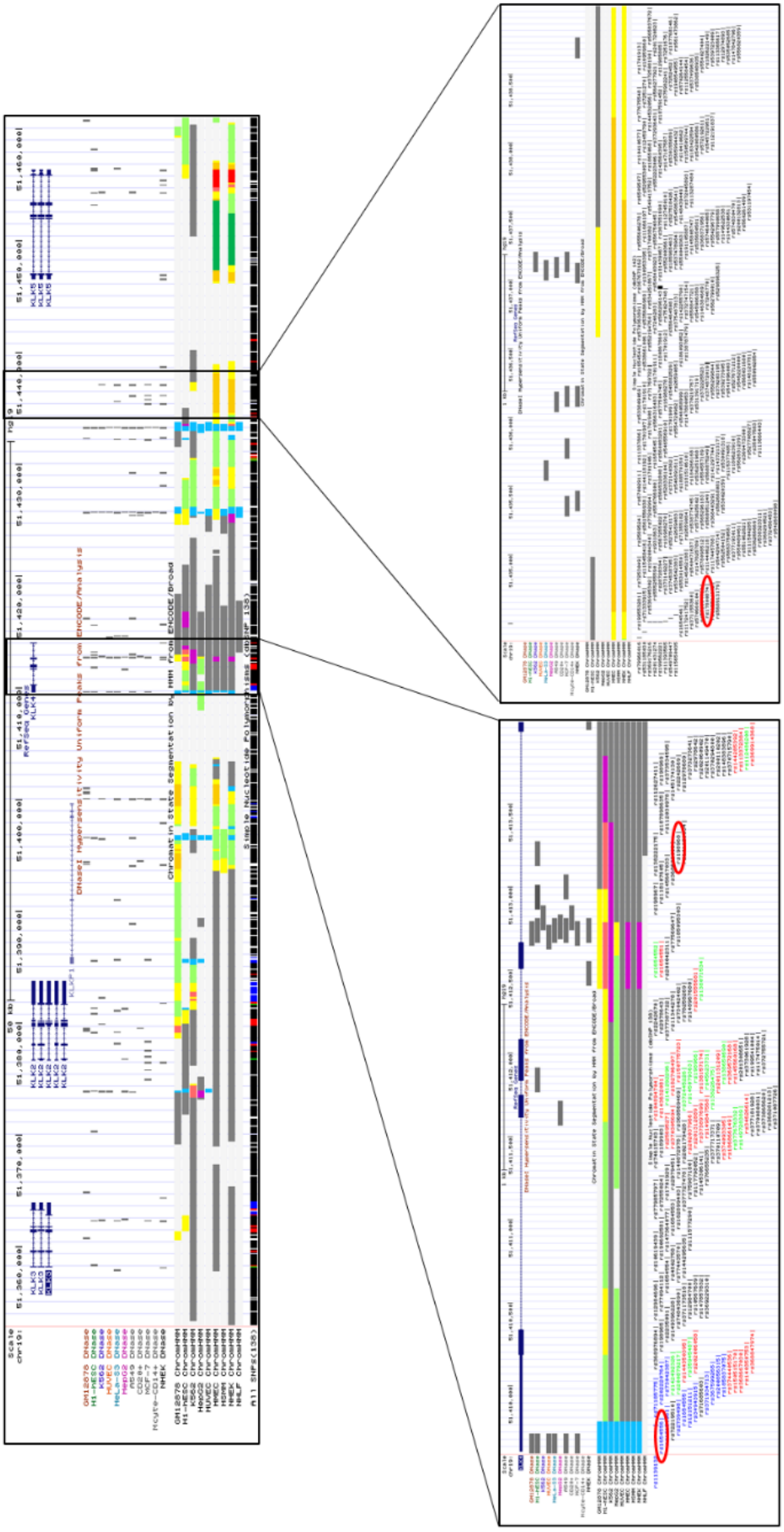


Figure S4 - Schematic representation of *KLK3-KLK5* landscape using UCSC Genome Browser. Reference genes, DNase hypersensitivity and chromatin state segmentation from ENCODE are shown in the upper image. The insets display in detail the *KLK4* locus and the putative enhancer within the intergenic region between *KLK4* and *KLK5*. The SNPs rs1654556, rs198968 and rs17800874 are highlighted by red circles.

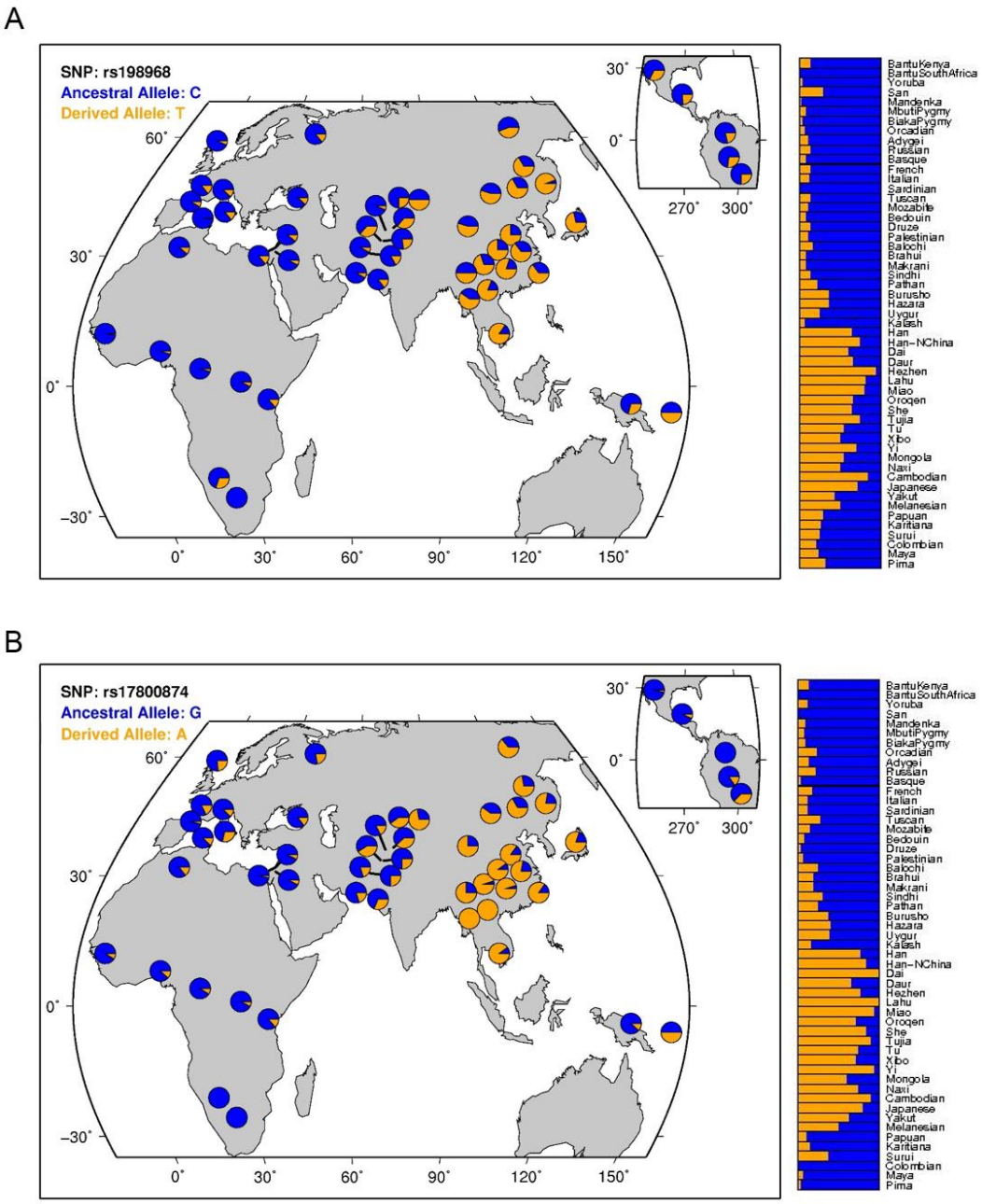


Figure S5 – Worldwide allele frequencies from HGDP data for rs198968 and 17800874 SNPs as inferred by fastPHASE (adapted from http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/)**. (A)** Frequencies of rs198968 located in intron I of *KLK4*. **(B)** Frequencies of rs17800874 located in a putative enhancer in the intergenic region between *KLK4* and *KLK5*.

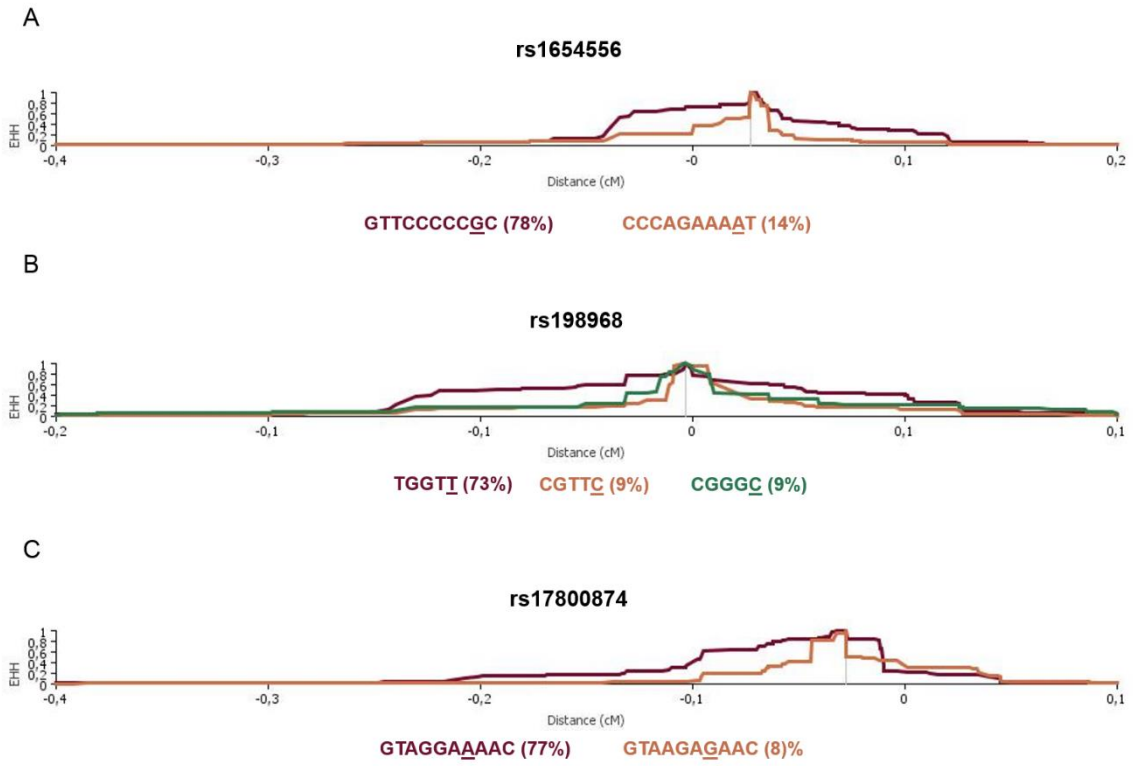


Figure S6 – Extended haplotype homozygosity (EHH) statistic for ASN (CHB+JPT) sample using 1000G data. Plots of EHH over genetic distance for the largest non-overlapping cores encompassing rs1654556 **(A)**, rs198968 **(B)** or rs17800874 **(C)** variants. Core haplotype sequences are indicated below EHH plots and candidate variants underlined.

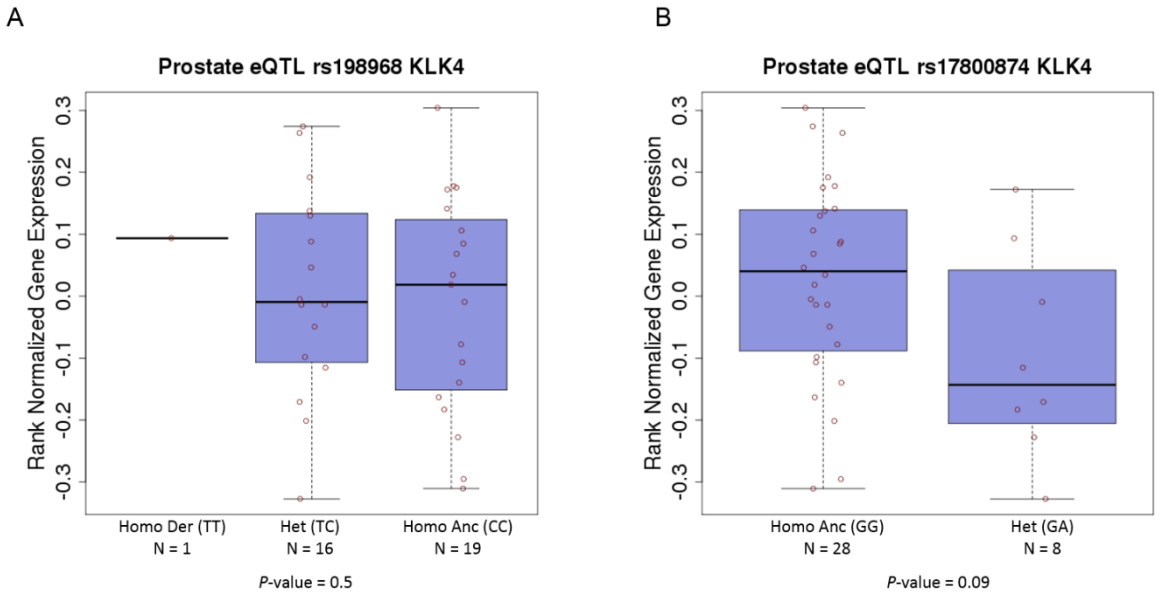


Figure S7 – Plots of *KLK4* expression for rs198968 (A) and rs17800874 (B) quantitative trait loci (eQTL) in prostate tissues from GTEx data (http://www.gtexportal.org/home/)**.** The corresponding genotypes are indicated in parenthesis and the number of samples and *P*-values are shown.

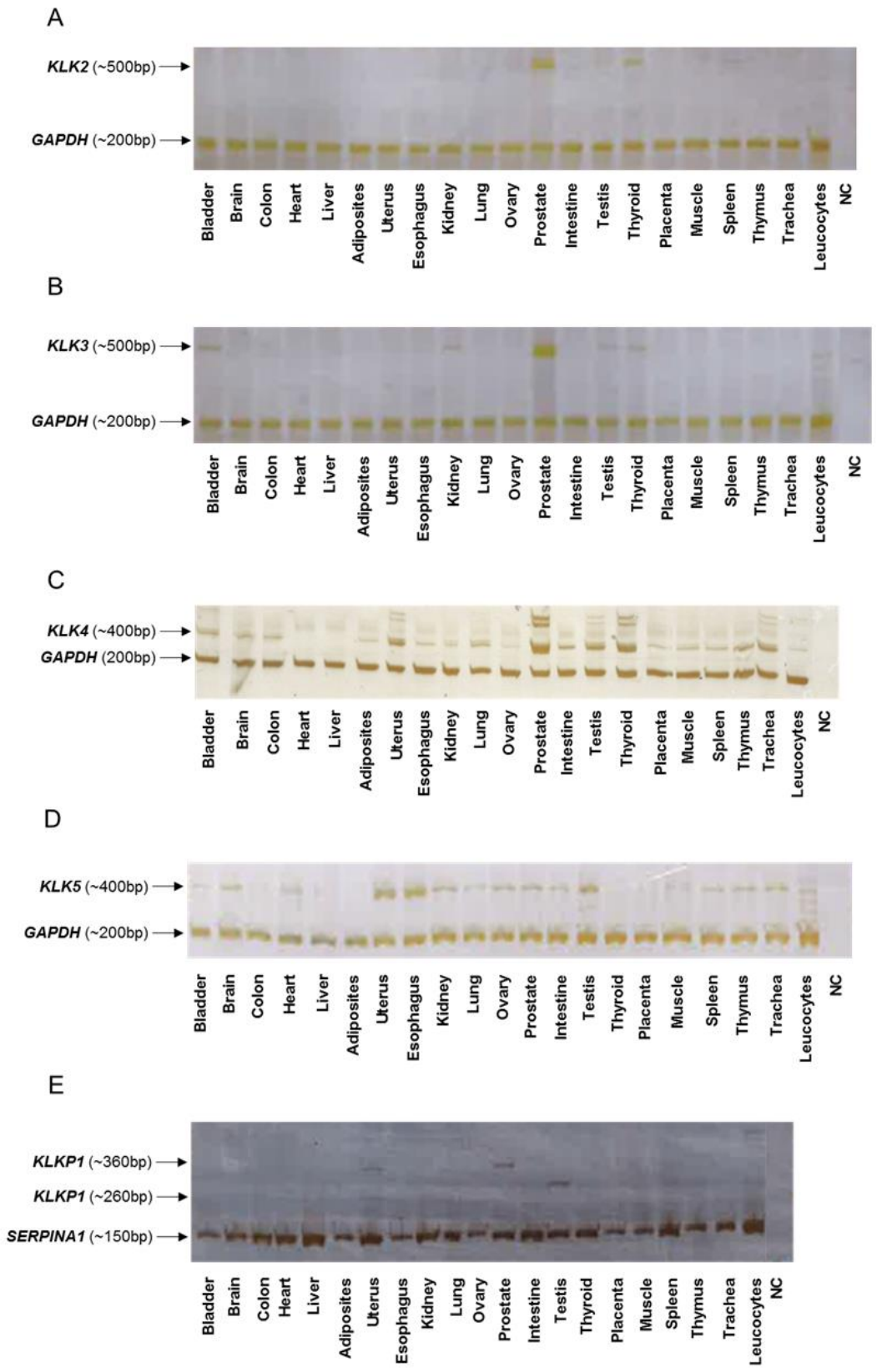


Figure S8 – Tissue expression of *KLK2, KLK3, KLK4, KLK5* genes and *KLKP1* pseudogene. Multiplex PCRs carried out in a cDNA panel from human healthy organs, each one including a minimum of three donor's pool. *GAPDH or SERPINA1* fragments were used as internal controls**.**

References

Gabriel SB, Schaffner SF, Nguyen H et al. 2002. The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Appendix C - Supplementary Material Paper III

Rare and common variants in *KLK* and *WFDC* gene families and their implications into semen hyperviscosity and other male infertility phenotypes

In preparation

Table S1 – Primers used for amplicon generation in pooled NGS sequencing.

Amplicon	Genomic Position (GRCh37)	Amplicon Size (bp)	Forward Primer Sequence	Reverse Primer Sequence
KLK1_1	19:51326900-51327089	189	GTGAGGCCAGCAAGAGAATC	TCTGAGGGGATAAGGGCTTT
KLK1_2	19:51324763-51325199	436	AGCAGAACCAGATCCCAAGAG	CTCCCTCTCCTAGCCTTGTCC
KLK1_2b	19:51326087-51326287	200	CAATGCAGGACCCACTTGAC	ACCACATGCTCCTTTCCTTG
KLK1_3	19:51323336-51323830	494	TTCCCTGGCCCTTTCTCC	TTCCTCTGTCTCCGCCCA
KLK1_4	19:51323064-51323400	336	AAGCAGATGCCTGGTTAGCTC	GGCAGACTGTGTAGCCCAAG
KLK1_5	19:51322376-51322718	342	TAGGTGATGGCAGAACGTGAC	CACCTTCCTCTGGGAGTGG
KLK10_1	19:51521652-51522148	496	AAGCGGTAAGGCAAGACTCA	AGCTGGGCCCTTCTTCTG
KLK10_1c	19:51522225-51522560	335	GTGACGGGAACACATTCTCC	CTGTTCCAGCCCGAATCTC
KLK10_1b	19:51522862-51523454	592	GTGGGGTGCAGGTAGCTT	GGCTCTTTTAGCCTTGGTTTG
KLK10_2	19:51520262-51520679	417	ACCTCCAGCTGTGGGAGTTC	ATATTTCCCTACCACCCTCGC
KLK10_3	19:51518996-51519504	508	ACAGACCCAGGCATCTAGGAC	ACTCTTCCTATCTCCAGGCCC
KLK10_4	19:51518499-51519011	512	GGATGGAAATGGGATTGAGG	AGATGCCTGGGTCTGTGAGTC
KLK10_5	19:51517931-51518318	387	TTTGAACAGTGCAGACAAGGG	TCGTCTTTATCCCAACCCAAC
KLK11_1	19:51530588-51530907	319	AAGGGAACAGAGCCCTTGG	CTGTCAGAACCTAGGCCCTCC
KLK11_2	19:51528768-51529183	415	TCACTGTCCAGACACAGAGGG	CCAACGACTTCCACATGGTT
KLK11_3	19:51527819-51528247	428	ATCCCTGAGCTTCTCTCCATC	GCCACTGCATTTGACCTTATG
KLK11_4	19:51527197-51527646	449	CACATCGTCACTGTGAACCG	TCCTCAAAGGTGTCACTCACC
KLK11_5	19:51526194-51526555	361	TTTGTGCTGAGGTCAGAAACG	TCACAGCAAGCACTGCATTAG
KLK11_6	19:51525706-51526066	360	AGGGTCTTGGCTTAGGGTTTC	AGTGCCTACCCATCCTGTCTC
KLK12_1	19:51537661-51537967	306	GAGACCGAGGTGAGCAGTACC	AGAGAAGCAGAGAGGGCAGAG
KLK12_1b	19:51538006-51538534	528	CTACCTGCTCCCCTGTGTGT	AACCCAGCTCCCTAGTCACC
KLK12_2	19:51537159-51537576	417	CACCCTCCCTTGTGATCCTAC	CCTGGGTGTCTTTGATGTTTC
KLK12_4	19:51533947-51534272	325	TTCCCTGATTCATCCTCCATC	GTATTGACATGGATGGGCCAC
KLK12_6	19:51532333-51532826	493	AAATGGACAGACATGGCCC	TAATGACCAGACTTGGAGCCC
KLK13_1	19:51567997-51568430	433	CAGCTTTCCAGAGAGGACAGG	GAGCAGAGAGGTTTCAGACGC
KLK13_2	19:51563600-51563974	374	TCCAACTCCATCCTCAAATCC	GCAATAATCCGACCCTTACCC
KLK13_2b	19:51566690-51567310	620	TTTTCAGAGCCTCAAGTCCA	GTTCTCTGGAAACCCCGTTC

KLK13_3	19:51563005-51563454	449	AGACCTTCCCTCCCATCTCTG	GGATGGAGAAGAAGTTAGGTTGG
KLK13_4	19:51561616-51562002	386	GAATTATTCCTGCCATCCCTG	TCGCCTATTTCACCTTCATCC
KLK13_5	19:51559725-51560128	403	TTTCAGGACATGGATCACTGG	TCTCCATCTTTCACTCTCTCCC
KLK14_1	19:51585668-51586143	475	AGGAATTTAGGCCTCCAGCAC	AGAGACACCTCCCCTCCTTC
KLK14_1b	19:51587364-51587653	289	ATACCTGGACCTGGAGCAGA	TTCCAACCCCGTCTGTTAAG
KLK14_2	19:51584672-51585120	448	AACCAGGAGCCAGCCACTAC	AATATTGGGCATGTCTCACCC
KLK14_3	19:51582504-51583087	583	TCAGACGTATGAGTCCAGGC	TGCAGTTAGTGAGAAGCAGGC
KLK14_5	19:51581193-51581509	316	GGACTCCTGGGTCCTGAGTA	CTGTCTCCCTGTGTCTGCTTC
KLK15_1	19:51334542-51334898	356	ACCTGACGGTAAGGGCTTAGG	GGGATACCGGTGGTCAGAAG
KLK15_1c	19:51335999-51336505	506	TATTCCTCCCACCGCATATC	CACGGGTTCAAGTCAGTTCA
KLK15_1b	19:51340287-51340546	259	TTCTGGAACCAGACAAGCTG	TCCCGTCTGCTCATGTGTTA
KLK15_2	19:51330807-51331153	346	TCGTCTTTCAGAACCCAGGAG	GTAGTTAGGGCAGGGTGCAAG
KLK15_3	19:51330044-51330489	445	CATAAATTCCGGCCCTTGTC	TCTTTCTGTCCTCTACTTGCGG
KLK15_4	19:51329759-51330100	341	TCCAAGCCTAACCCTAGCATC	CAAGGAATCCTATGCTCCAGG
KLK15_5	19:51328912-51329337	425	TGGGACAAGTCCTTGGCTAAC	ATTGAGTTGAGTTGGGTTGGG
KLK2_1	19:51364665-51365253	588	TAGGGGAAGGTTGAGGAAGG	CCTCACCCCCTGTACAGAAA
KLK2_1b	19:51376610-51376917	307	TCAGCATCTAGGTGCCAACAG	TTGCAATGTGGAAGAGTCAGG
KLK2_1c	19:51378707-51379223	516	GCTGTAGCTGGGTGCACAAT	GAGGAGGAGAGGCATGAGG
KLK2_2	19:51377818-51378341	523	TAGCTACAGAATTGCCAGCCC	CGAGATCAGACACAGAGAGAAA
KLK2_3	19:51379646-51380135	489	TGGAGTCTCCCTTATCCTCCC	GCGCAAGACTATGGGTCAGAG
KLK2_3c	19:51380421-51381106	685	TTCCCTGACTCCCTCAACAC	CAGGTCATCCTCATCGTCAC
KLK2_4	19:51380027-51380360	333	CAGATGGTGTAGCTGGGAGC	GGGAGGCAGATGTCTGGTTAG
KLK2_5	19:51381548-51381890	342	AGAGCTGGGAATTGCTCTCAG	ATCCAGAAAGGCCAAGTGATG
KLK3_1	19:51358007-51358329	322	GCAAGTGCTAGCTCTCCCTCC	GGGAAAGAGCCTCAGCTTGAC
KLK3_2	19:51359371-51359744	373	GACTCCCAGCCTTGGTTCTC	GGCCTTAGAGGTTATCCTGGG
KLK3_3	19:51361204-51361670	466	GAGCCTCCTCCTTGCTCC	GTTCCTTGGCCCTCCTCC
KLK3_4	19:51361552-51361998	446	CAGCATTGAACCAGAGGAGTG	AGGGAATGAGATGAGACACGG
KLK3_5	19:51363154-51363496	342	GGTCTGAAAGATAGGATTGCCC	ACCCTGGACCTCACACCTAAG
KLK3_5b	19:51362765-51363091	326	CCATTCTCCACCTACCCACA	CCGACTTCCAGAAGAAGGTG

KLK4_1	19:51413768-51414071	303	CACGATACAAGGAGTTGCAGG	GAGAGGGATGGAGAGACTTGG
KLK4_2	19:51409576-51410163	587	GGGAAGCAAGGAGGACACTA	TGACCCCCAAATACATCCTG
KLK4_2b	19:51412404-51412768	364	AGAGCCTTCACCGCTGTTTC	CTCCTGAACCTCTGACCACG
KLK4_3	19:51411022-51411567	545	CACTGTTTCCCTCTGGGTACA	AAGGGGGAGACAGAGACACA
KLK4_3b	19:51411761-51412422	661	AGAGCTCTGGGTCAGCCC	AAACAGCGGTGAAGGCTCT
KLK4_4	19:51411514-51411823	309	TCTCTCCATCTCTGCATCTCG	GGTGTGTGTCTGCCCTCTTC
KLK4_5	19:51410044-51410500	456	GGGATCTGTACCCTTGGTTTG	ATCTGGAATGGGACTTCCAAC
KLK5_1	19:51455688-51456387	699	GATCGCACAACCACAAGTACA	GCTATTGCTAAGGCCCGATA
KLK5_2	19:51452869-51453482	613	CAACCTCATCCTCCCACCTT	TCAGCATGTCAGACACCCAC
KLK5_4	19:51451819-51452504	685	CTCAGAATTTGGCAACGCTC	GAGGGAGTTGAGGATGGTTTG
KLK5_5	19:51446793-51447129	336	TCAACATCTCTGGGAAGGAATG	CTATGGGCATCTCTGGGTCTC
KLK6_1	19:51471244-51471553	309	CCCAATACCAGCCTCTTCTCC	GGGATCCTCTGATGGAAGATG
KLK6_E2	19:51471710-51472089	379	AGCCAGTCTCCTGGCTCAGG	ATGGGAGTTTTCCTCGGAGC
KLK6_1b	19:51472706-51473027	321	AGCCTGCCCAGGTTCAGT	GGACAAAAGGAAGCCATTGA
KLK6_2	19:51470337-51470659	322	GGGATGCCTATGTCACCTCC	TTGACTGGAGTTCATGTTGAGG
KLK6_3	19:51466444-51466886	442	CTCACCATTAGCCCATCTTCC	ATTCATGACTTCCCAGCCCTC
KLK6_4	19:51464861-51465166	305	GGGTTCTCCCTCAGCCTGT	GCTGAGTCTGGCCCATCTCT
KLK6_5	19:51462288-51462771	483	CACGTCGCTGCGTTTATTAAG	CTGTGCCTTTGTGTGTCTTACTG
KLK7_1	19:51485315-51485741	426	AGACAGACAGAAGGAGCCCAC	CTGGGCCTGAATGCTTTCTC
KLK7_1b	19:51486996-51487408	412	TCAGGACCGGGAGTCTAGG	GGACTGTGGGACCAGAATGT
KLK7_2	19:51484926-51485262	336	GAGAATAGAGCCGTAGGCACAG	CTGTCCATCTCTGACTCTGGG
KLK7_3	19:51483232-51483772	540	CCAGAAGGGCTGTTGTTCAG	TCTGAATCTGTGAATGTCACTCC
KLK7_4	19:51482961-51483262	301	CAGTCACTCGGGTCAAGCTC	GAACCAGGAGCCTGAACAAC
KLK7_5	19:51480716-51481057	341	ACTTCATAGGTCATCGGCGTC	TGGTGGAAGTCCATAATCTGC
KLK8_1	19:51504766-51505458	692	TCTGAGTACCTCTGCACCTCA	TTCTTCGGTTCCCGGTTACT
KLK9_1b	19:51504228-51504595	367	CCGTATCGCACTCTTCACATC	CCTCCCTTGGAATGTAGGAATC
KLK9_2	19:51503604-51503946	342	TCTCCTTCCTGAGTCGAAACC	GTTGTATGCGGAGAACTTGCC
KLK9_2II	19:51512357-51512986	629	ACTTCTGGCCTAAAGCACCC	GCACTTACCTCCTCTCCCAAG
KLK9_3	19:51503178-51503608	430	CTTTGCCAATAGCTTGGGTTC	GGAGAGAGCTGAGGACTGGAC

KLK9_3I	19:51509588-51510089	501	TCGATTGGCAGATAGATAGACTGAC	GTGTGGCTTGCAGAGGGTC
KLK9_4	19:51500783-51501212	429	AACAGCAGCAGCAATTCAATG	GGGTTAGATCCAGACAGGGATG
KLK9_5	19:51499160-51499568	408	CCATGAGAACTTGGTTTCTGC	ACCAATCATGCCAAAGAACTG
KLK9_5I	19:51506182-51506643	461	ACTGTGTCTTGCTTGACCTCG	GAGATGGAGGACTCTCGGATG
KLKP1_1	19:51399490-51399948	458	TCCCTCCGATACCTCCTCTT	TAGTCATGCCCTGCTCACTG
KLKP1_2	19:51398587-51399086	499	GGAAGGTGTGGGTTGTATTGA	ATACCCTGCCCATGTTTTCC
KLKP1_3	19:51398300-51398561	261	AGACCCACCCAGCACTCA	TGAACCCCTCCACTTTGTGT
KLKP1_4	19:51390924-51391588	664	TATTCCTGTGGGGCACAGAC	CCCTGATCTGGGTGTAGCAT
KLK-prom51323816	19:51323816-51324326	510	GCGGAGACAGAGGAAGAAAG	CCTCGACCTTGGACTTCAGG
KLK-prom51324011	19:51324011-51324670	659	AAGACAGGGACAGCGAGAAA	AGTCTCTCTGCGGCATCTCT
KLK-prom51326564	19:51326564-51327093	529	CAGGAGGGGATGATCAGAGT	GGGCTCTGAGGGGATAAGG
KLK-prom51339854	19:51339854-51340550	696	CCCCAGCCTTGTCCCTTC	GCACTCCCGTCTGCTCAT
KLK-prom51361995	19:51361995-51362670	675	CCCTCCTTCCCTCTTCTTTG	CCATCATCACTCCCTCCACA
KLK-prom51373385	19:51373385-51373812	427	AGGAGGAATGTGGGTTCTGA	TGCTGAATGATGAGTGGATGA
KLK-prom51373661	19:51373661-51374136	475	GTTGCCCATGCTTTGATCTT	TGCCTCGTATCTGGGAGACT
KLK-prom51376037	19:51376037-51376736	699	GGCACATGAGACTTTGTATTGAA	CCCACATGCTGACACAGG
KLK-prom51393934	19:51393934-51394591	657	CACACATGTGAGCCATGTCC	GTAGCCCCTCCAAAAGCAAT
KLK-prom51398232	19:51398232-51398523	291	GCCTGGGACTCTCCTTCACT	AGCTGTGATTCCCCCTGAAG
KLK-prom51412383	19:51412383-51413066	683	GGCTCATTCCGTCCTCCT	CTACCCTGAATCCCTGACCA
KLK-prom51424940	19:51424940-51425412	472	TGGCTCCTCTGGTTATGGAG	CCTTTTGTGGTGTTGTTTCG
KLK-prom51425335	19:51425335-51425815	480	CATGGACACACAGAGACATGG	TTCCTGCCACCTCTGCTC
KLK-prom51456150	19:51456150-51456555	405	GCACAGACACCTCTCCTTCC	TGACCCAGAGTTGGTGAGAA
KLK-prom51471042	19:51471042-51471502	460	GGCTCTCTCCAACTTTCCAA	GATGGGAAGGACAGAGGTCA
KLK-prom51472627	19:51472627-51473027	400	CACCCCCAGCACTCTCTG	GGACAAAAGGAAGCCATTGA
KLK-prom51482248	19:51482248-51482852	604	AAGGGAGGGAGAGCAAGCTA	CCTGGGGATGAGACAGAGAG
KLK-prom51495240	19:51495240-51495641	401	CCTGGCACCTGTTGCTAAGT	CCCAAAAGAAATGGGGTTCT
KLK-prom51495933	19:51495933-51496293	360	CCCCAAGTCTTTTCAGGCTTA	AAAGAGACTCGTTTTAAACATAACCAA
KLK-prom51504253	19:51504253-51504749	496	CACACGCACCCACATAACC	CGGACCCTCCTTCTCCAG
KLK-prom51505824	19:51505824-51506294	470	CAGACCCTCCCTCAATTTCC	GTCCCAGCCTCAATGGTTC

KLK-prom51522289	19:51522289-51522745	456	CCGACCTTACCCCAGAGTTG	TGAGAAAGAGGCTCCCACTG
KLK-prom51522567	19:51522567-51523264	697	TGGGCACAATTACCCTAATGA	CTTGCTGGGGACGTGAAC
KLK-prom51529766	19:51529766-51530361	595	GGCAGCTTCCCTTTCCTC	GAGGGATTTCAAGACAAATTGC
KLK-prom51530756	19:51530756-51531088	332	CCACCTCAACCTCTGCATCT	CTTTCCAAGTGGACCCTCCT
KLK-prom51569993	19:51569993-51570442	449	GGGCTCACATGGAAACTTGT	TGTCGCCAATGCAAAGTTAG
WFDC5	20:43738985-43739464	479	ATTAGAGCTGGGCTCGGAGA	ATAACCGTGAGCTTCCTCCA
WFDC5-I	20:43743619-43743832	213	AGAGGGATTTCCCGCACT	GTCACTCCACACCCCCTCT
WFDC12	20:43752727-43753154	427	CTGGGAGGCAGGTCTCCTTA	ATGCCCAGCAGGAACACTAT
WFDC12_3	20:43752037-43752609	572	GTCTGAACATCTCCATTGTCCAAAG	GCTGCTATAACAAATTACCATAGACTGG
PI3	20:43803421-43803692	271	CCCAGAATGGGGTGGATATT	AGGGTCTCCCCTTAGTCCAA
PI3-I	20:43804476-43804801	325	TGTGGGCTCGTTTCTTCTTT	CCTCGTTCTCCAGCTAGTGC
PI3-II	20:43804954-43805220	266	TGCCTCTGAGTGCTTTGATG	CCAGGAGCCCAGAAGTCATA
SEMG1	20:43835618-43835803	185	CACCCATGGCACACTCACT	CAGCTTTCCCCAAGGCTTAC
SEMG1_E2_1	20:43835886-43836527	641	GGAGAACATAACTGCTTTGGGATCC	GATTTTCTCCACCCAAAGCTTCAG
SEMG1_E2_2	20:43836371-43837103	732	GGATTTGAGACTTGCCTGCTA	GATCAGGGGAATAGCCCATC
SEMG1-I	20:43838184-43838456	272	ATCCCCACCCTTCACTTTTT	CCCCTTGAGCCACTAGGAAT
SEMG2	20:43849891-43850127	236	GGTGAGGAAGCTGGCATTTA	AGGCTTACCCTCTCCACTCA
SEMG2-I	20:43850295-43850958	663	TTGCAAGAGAGCTTTGGAGA	CCTCTCTCACGTCAACCACA
SEMG2-II	20:43850835-43851528	693	GGTGGATCCCAAAGCAGTTA	TGAGACGTCTTTCTTCTGTACTCG
SEMG2_E2_1	20:43851192-43851925	733	CTGGATTCCTGTTTGTATCTGCCTT	AAAGATGTATCCAAAGGCAGCATTT
SEMG2_E2_2	20:43851683-43852219	536	TCTACTCGCCAGGAAATGGTGTTCA	TTCAAGTACAGAAGAAAGACGACTC
SEMG2-IIII	20:43852881-43853120	239	ACCCCCACTCTCCACTATCC	ATCACTGGAGCCAAAAGCAG
SLPI_1	20:43882078-43882825	747	TCTTCACCTGTCTGCGACTCT	GCTACCTTCCTGCTTAGGGC
SLPI_2	20:43881536-43882286	750	AGGGAAGAAGAGATGTTGTCCTGAC	TGCTTGGAGTAGGGTTCAGG
SLPI_3	20:43880699-43881240	541	GCTTTGAGTTTAGAGTTTTACGGTG	AAATCACTTGTCCCAAATCACAG
WFDC2-I	20:44098945-44099303	358	GGTTAAGGTTTGGAGCAGGA	TTTCAACCGCCTTGACTTTC
WFDC2-II	20:44099770-44100098	328	AGGGTTCAGCTTTTGCCTTT	CACAGGGGCTGTATTCTGGA
WFDC2-III	20:44109941-44110192	251	GCTGCAGGTACTCCTGCCTA	GGGGAGACAGAGACAGAACG
SPINT3	20:44141031-44141518	487	TTGCTCTAAATTTCCCCAGTG	GGTTTGGTCTCCTTCAACCA

SPINT3-I	20:44144124-44144307	183	CCAGCCTGATTCCCAGTCTA	TTGGTCCCTTTTCTCCCTAAA
WFDC6	20:44162813-44163181	368	GAATTATTCAGCCTCGGAGAG	GTCCAAGCATCTCACCCTTT
WFDC6-I	20:44163762-44164007	245	CTTGGTGAGCACAGGAGACC	CAGGCACCCAAACAGCTCTA
WFDC6-II	20:44167931-44168167	236	CAAGGACGGAGTTCCCAATA	TGTCTCCAGGGCTGTCTCTT
EPPIN-WFDC6	20:44165605-44166266	661	ACAGATGCCTGGCTGGAA	TGCCAACATTGTGTCAGGTT
EPPIN-WFDC6-I	20:44166557-44166786	229	GATAGGAGGTGAGGGCACTG	GGGAGCACTGGCTCCTTTA
EPPIN-WFDC6-II	20:44170519-44171204	685	TGACACAGGTATAAACTGGGAGT	AAGAATGCACCTGGCAAAAG
EPPIN-WFDC6-III	20:44175884-44176435	551	AAGGGCTGAGGCCAATACTT	CCAAGAGACCGACAGAGACA
EPPIN_E2	20:44174141-44174883	742	ATTCCTCAACAAGGCCAAAAAGA	TAAATCGCAGGTCTCCCAAGTC
EPPIN_E3	20:44171201-44171700	499	AGGGATGCAGATGCCCACTC	TCTTACCTCCTGATCAGAGCTGG
WFDC8	20:44181751-44181982	231	GGGAGGTATATCCCCAATCC	AAGGCTCTTTGAAGGGTGGT
WFDC8-I	20:44184309-44184529	220	GACCCTCAGATTTCCAGGAT	TTTCCTTGCTTCCTGACCTC
WFDC8-II	20:44187453-44187668	215	AAGTGGGGAACCTCTGTCCT	CCGAAAGAATGTGAGTTTGGA
WFDC8-III	20:44190678-44190883	205	GGCCTCATGTCTAACCCTCA	CCCACCTCTCTTCTCATCCA
WFDC8-IIII	20:44207815-44208012	197	AGACCCTCCATGTCTGGCTA	GGGCAGACTGGCATAGAATG
WFDC8_E6	20:44180310-44180901	591	GAATAAAACCTTGTTTGAGATTCCC	TGCAATAAAACTTATGTATTCCAAA
WFDC9	20:44236511-44236818	307	CTGCTGAGGATTGCAATGAG	CCCAGGATTCTTGACTTTTGA
WFDC9-I	20:44237241-44237477	236	TCGGTAAGGACCAGTGATGA	CGGACCCTGCTAACTTCTTC
WFDC9-II	20:44238707-44238915	208	CCGTCCCTTTTCACACCAC	ACCCTGTTCTCCTCAATCCA
WFDC9-III	20:44243211-44243406	195	AATCCAAGCAAGCCACAAAT	AAGTTCCTGGCCTTTGAACA
WFDC10A	20:44258078-44258562	484	TGCAATGAAACATCAGTTAATTCT	CCAGCAGCCCATCACCTAC
WFDC10A-I	20:44259437-44259927	490	GAGGAGGCTTCAGCTTGAGA	ATTGAACACCAGGACGCAAG
WFDC11	20:44277155-44277451	296	AAGGGTACTCATGGGTGGTG	TTAAGCAGGGGGAATGAATG
WFDC11-I	20:44277868-44278072	204	ACACCCAAGCAAACATCTCC	CCATCAACCTTTGACCATCTC
WFDC11-II	20:44279118-44279335	217	CTAATCCAGCCTCACCACCT	TTGCTTCACACCTCTCTTTCC
WFDC11-III	20:44295650-44295844	194	AACCCATCCAAGCAAGACAA	TTGACTGGTTTTGCTCACCA
WFDC10B	20:44313261-44313619	358	CACCTGTAGTGGCAGCAAAA	CTGCTCCTGCATCATTTCAG
WFDC10B-I	20:44314506-44314715	209	ACTCTCCATCACCCATCCAG	ATTTCCTTCCCTCTCCACAG
WFDC13	20:44330630-44330869	239	TTGGAGTACACGGTGAAAGG	GCCCAGATCCAGCACCTAC

WFDC13-I	20:44333046-44333680	634	TTGGTGAGGCCCTTCTCTTA	ACTTCCTTTGGGCTGGATCT
WFDC13-II	20:44334363-44334590	227	TTGAGGTTCTGCCTTCCCTA	GCGAGACCAAATCCCTGTACT
WFDC13-III	20:44336383-44337027	644	GCCTGACAACAGACAGTTCAA	GCACCTTATTGCAGCCTCTT
WFDC13-IIII	20:44336940-44337566	626	GCCATATCATGAACCCAGGA	TGGAAGGAATGGAGGCAAG
SPINT4	20:44350923-44351157	234	GGTGAGCTCAACCGTCTCA	TGACCTTTGGATTTATTTTCTGC
SPINT4-I	20:44352497-44352727	230	TCAATGGGAACCTCTCATGC	CAAGGACCAAGGACAGGATT
SPINT4-II	20:44354240-44354542	302	TGGGTTTTTCTTTCCTTTGC	GATGGGACCAGCTGACATTT
WFDC3	20:44386959-44387160	201	ATTGCCCTGCTACATGCTTC	GGGCCTGAGAGTTCTCACCT
WFDC3-I	20:44402824-44403163	339	TGAGCCCTCAAATAGCACAA	TCCTAGGGTGGAGGGAAAAC
WFDC3-II	20:44404022-44404269	247	TCTGTTATTGTTGATGACAATGGA	CCTTTTCCTGTGCCTGTTGT
WFDC3-III	20:44405690-44405873	183	TTCTTCCCAAAAGAGCCAAC	GGGCTCCAATGCTTTGATTA
WFDC3-IIII	20:44408377-44408577	200	CGTCCCAGATTTCACCATTT	GTGTCAGACTGGGTGGCATT
WFDC3-IIIII	20:44416429-44416645	216	CAAGCAGACCAAACAGACCA	TTGCTGTCTCCCTTTTTGCT
WFDC3-IIIIII	20:44416626-44417321	695	AGCAAAAAGGGAGACAGCAA	AGGCTGGAGGGAAGAGTTGT
WFDC3-IIIIIII	20:44417045-44417736	691	CCTCTCAACCCTGGTGCTAA	TGGCTAGTTACCCCAAACTCA
WFDC3-IIIIIIII	20:44418509-44418733	224	CAGGGATCCCAAATGATAGC	AAGGTTTTGGGACTGGGTTT
WFDC3-IIIIIIIII	20:44419466-44419660	194	GTCCCTCTAACCCACACAGG	GCCAGTGGACACCATTAGGT
WFDC3-IIIIIIIII	20:44420406-44420624	218	AACACCCACACTGCACAATC	ACTCCGCTGGACTCTGTCAC
promWFDC43740354	20:43740354-43740863	509	TGCACAAAGCTGGAAATCAG	AGGGCTAAACCTTGGACCTC
promWFDC43743055	20:43743055-43743675	620	AGAGCTTCCTCCCCACAAA	GTCAGCTGCCTGCTGTCTTT
promWFDC43743631	20:43743631-43744251	620	CGCACTCACCTCCCTTCTT	GCTGGCCTCACTCATTATGC
promWFDC43756560	20:43756560-43757240	680	GGTAAAATTGCAGGGGACAA	TGGTTCAAGGCTTTCCACTT
promWFDC43767320	20:43767320-43767859	539	AGAAAGATGCCACAGGGTGA	TGATACAAGCCTGGGGAAAG
promWFDC43803065	20:43803065-43803761	696	CAGGACCAGGGAAGAAGGA	CTGAAATCTGGGCTCCAGTG
promWFDC43805135	20:43805135-43805514	379	CTGGAGCTGCCTCTCTCATC	AGGACTGATGGTGGATTGGA
promWFDC43810346	20:43810346-43810687	341	GGCAACTTTTGGGAGGAAGT	GGCATGAGAAGTGGCTTGAT
promWFDC43835443	20:43835443-43835710	267	TGGCATGATGATCTAAAAGGA	TGATGTTGGGCTTCATCTTG
promWFDC43838631	20:43838631-43838893	262	AGGCTTTCATGACCCTGGTA	CTTTCCATCCACAGGTGACA
promWFDC43847932	20:43847932-43848286	354	TGCATTTTCCTGCTTTGTTTT	TGAAGCATGTGAGCTATGTGG

promWFDC43880574	20:43880574-43880886	312	TGCATAAACTCGTCTCAGGAAA	CTTTGCACATCCTGCTTCTG
promWFDC43881954	20:43881954-43882216	262	GCACCTGGCTCTCCTAGAAC	AGTAAGCAGGTCGGGGAACT
promWFDC43882809	20:43882809-43883391	582	TCGCAGACAGGTGAAGAAGA	GTTGCTGTGTTGGCCTCATA
promWFDC43883326	20:43883326-43883925	599	TCCTGACACCAAGGAGATTG	TCCCTTCATGCCCTTAATTTC
promWFDC44096513	20:44096513-44096955	442	TGATGCCATTGACTGATGAAA	GCCACTGTCCCCAAAATG
promWFDC44118145	20:44118145-44118600	455	GGGTGAGGGAGAAGGAGAAT	GGTAGTCACACACCGAGTTCC
promWFDC44126775	20:44126775-44127304	529	TCCCTGGATGAACAATCCATA	CTATCATGAGGCTGCAGAGC
promWFDC44144077	20:44144077-44144479	402	AAATGGGCTCTGTTTGTGCT	GCTTCCACCCTGAAGTGGTA
promWFDC44146706	20:44146706-44147345	639	GGAAGTGAGAAGCTGCCTTT	TGATTTTTCACATACATGCACA
promWFDC44166833	20:44166833-44167497	664	CCACTTCACCTATCCCCAGA	CAGCTAACTGTACTGGGCACTG
promWFDC44217442	20:44217442-44217785	343	CCAGGAGGGTGTTTCTCACT	GTGGAGAGGTGAAGGTGGAA
promWFDC44239700	20:44239700-44240140	440	GAACTGCAAATGGCCAGAC	TTCCTTGCTTCTTCAGGAAA
promWFDC44258078	20:44258078-44258756	678	TGCAATGAAACATCAGTTAATTCT	TGCAGGTTTCTTCTTCATTCC
promWFDC44314543	20:44314543-44315008	465	TGCATCCTCATCTTGTCACG	TGGGTTAATTCTAAGCACCTTTT
promWFDC44333893	20:44333893-44334284	391	TTGCAAGGCCCATATGAAAT	GCAAGGGCTAGTCTCTTCAAA
promWFDC44386629	20:44386629-44387010	381	GCCAGAGACAGGTAAGCAGAA	CCCTGAGGGTTTCCTTCTTT
promWFDC44407439	20:44407439-44408065	626	TGTGGCATCTAACCATCGTC	CCTGGGTTGTTTAGCTGCAT
promWFDC44409920	20:44409920-44410617	697	CATCATGATCCCCTACACTGC	GGCTGACCCTTTTCTAAAACC
promWFDC44420732	20:44420732-44421329	597	GGCAGCTGGTTCTTACGG	ATGTTCCAAGGGTTCTGTGG

Table S2 – Variants identified in phase I.

Due to the large size of the table this data is provided in digital format (Supplementary Tables – Paper III).

Table S3 – List of common SNVs identified in phase I and presenting significant nominal *P*-values for case-control association analysis by Fisher's exact test.

Gene	SNV ID	Consequence	Control *vs.* Cases (HV+NV)	Controls *vs.* HV Cases	Controls *vs.* NV Cases
			***P*-value**	***P*-value**	***P*-value**
upstream *WFDC12*	rs6104024		0.0010	0.0278	0.0027
upstream *WFDC12*	rs6017495		0.0251	0.0441	0.0851
intergenic	rs6073760		0.0468	0.0630	0.1475
WFDC6	rs41304411	p.I122I	0.4500	0.0002*	0.0516
intergenic	rs6073789		0.1218	0.6906	0.0493
intergenic	rs6065839		0.0840	0.1988	0.0338
upstream *KLK15*	rs266851		0.0781	0.2009	0.0305
KLK3	rs2271095	intronic	0.0391	0.2129	0.0349
KLK3	rs266875	intronic	0.0189	0.2012	0.0346
KLK3	rs11084034	intronic	0.0410	0.3432	0.0415
KLKP1	rs2739491		0.0813	0.1762	0.0386
KLK4	rs2979451	intronic	0.0863	0.0246	0.1346
KLK6	rs28384475	5'UTR	0.0607	0.1204	0.0386
KLK7	rs1654526	intronic	0.0006	0.0003*	0.0066
KLK7	rs1722558	intronic	0.2688	0.0780	0.0090
KLK7	rs1991820	intronic	0.1548	0.0120	0.2560
KLK7	rs1991819	intronic	0.2781	0.0255	0.4339
KLK7	rs1991818	intronic	0.3778	0.0470	0.4756
KLK10	rs2075691	intronic	0.0461	0.1990	0.0896
KLK10	rs2075690	p.L149P	0.0320	0.1408	0.0729
KLK10	rs2075687	intronic	0.0306	0.2756	0.0468
KLK10	rs77303625	intronic	0.1544	0.0278	0.3547
KLK10	rs10425377	intronic	0.0623	0.0073	0.1895
KLK10	rs7259651	5'UTR	0.1139	0.0152	0.2616
KLK12	rs75565227	intronic	0.1336	0.0398	0.2717
upstream *KLK12*	rs8104577		0.0230	0.0294	0.1234

KLK13	rs2736433	intronic	0.0487	0.2756	0.1532
KLK14	rs11671800	intronic	0.0188	0.0269	0.0826
KLK14	rs6509518	intronic	0.0031	0.0225	0.0157

Significant nominal *P*-values ($P < 0.05$) are underlined.

*Significant *P*- value after adjustment for multiple testing (Bonferroni correction).

Table S4 – Identified SNVs in *SEMG1* and *SEMG2* in the pilot study.

Gene	SNP ID	Consequence	Control	Cases (HV+NV)		HV cases		NV cases	
			MAF	MAF	*P*-value	MAF	*P*-value	MAF	*P*-value
SEMG1	rs17850164	T293T	0.019	0.000	0.0423	0.000	0.1287	0.000	0.1484
SEMG1	rs2233887	R372L	0.013	0.003	0.2893	0.007	0.5195	0.000	0.2880
SMEG1	rs147894843	G400D*	0.000	0.010	0.2663	0.013	0.2364	0.007	0.4626
SEMG1	rs7270676	N402N	0.051	0.024	0.7107	0.027	0.2152	0.022	0.1642
SEMG1	rs79500955	R447H	0.006	0.007	0.7120	0.007	0.7377	0.007	0.7120
SEMG1	rs2233889	N455N	0.019	0.028	0.4077	0.020	0.6330	0.037	0.2823
SEMG1	5 repeat units		0.051	0.024	0.1190	0.027	0.2152	0.022	0.1642
SEMG2	rs2233901	S274N	0.139	0.178	0.1765	0.193	0.1307	0.162	0.3528
SEMG2	rs2233903	H279Y*†	0.019	0.028	0.4077	0.020	0.6330	0.037	0.2823
SEMG2	rs2071650	G368R*†	0.019	0.028	0.4077	0.020	0.6330	0.037	0.2823
SEMG2	rs139977707	E552Q*	0.000	0.010	0.2663	0.020	0.1143	0.000	N/A

*Predicted as damaging by SIFT or Polyphen.

†Variants in linkage disequilibrium

N/A - Non-applicable

Significant nominal *P*-values ($P < 0.05$) are underlined.

Table S5 – Low-frequency burden analysis of *SEMGs* predicted functional variants.

		Cases (HV+NV) *vs.* Controls	HV Cases *vs.* Controls	NV Cases *vs.* Controls
Nonsynonymous	Total	C-alpha = -2.9879 (*P*-value = 0.6566)	C-alpha = -1.8499 (*P*-value =0.6180)	C-alpha = -0.6938 (*P*-value = 0.8992)
	SEMG1	C-alpha = -0.0493 (*P*-value = 0.8195)	C-alpha = -0.4830 (*P*-value = 0.6813)	C-alpha = -0.0926 (*P*-value = 0.8670)
	SEMG2	C-alpha = -2.9186 (*P*-value = 0.6933)	C-alpha = -1.3669 (*P*-value = 0.6738)	C-alpha = -0.6011 (*P*-value = 0.8194)
Damaging substitutions	Total	C-alpha = -2.4665 (*P*-value = 0.7788)	C-alpha = -0.8140 (*P*-value =0.8282)	C-alpha = -0.5609 (*P*-value = 0.9173)
	SEMG2	C-alpha = -2.9186 (*P*-value = 0.6935)	C-alpha = -1.3669 (*P*-value = 0.6823)	C-alpha = -0.6011 (*P*-value = 0.8197)

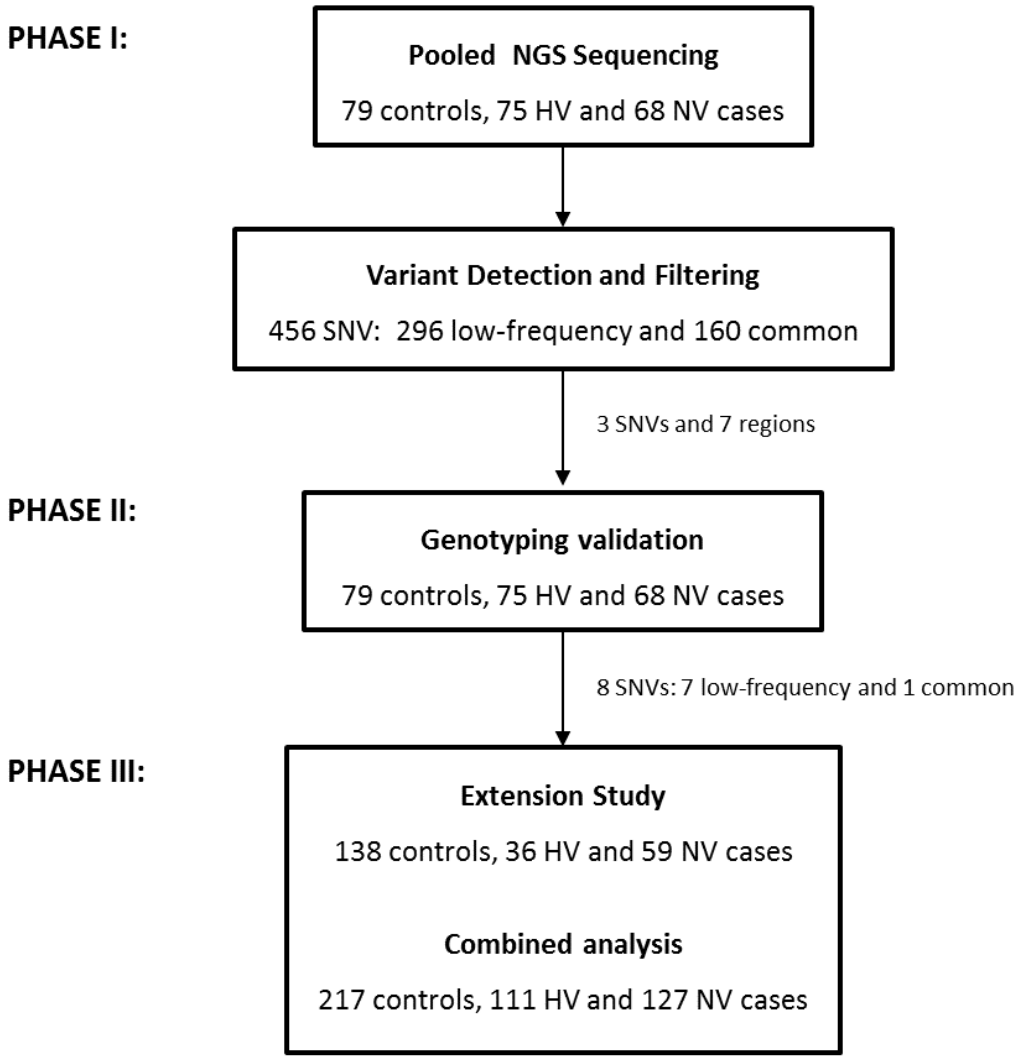
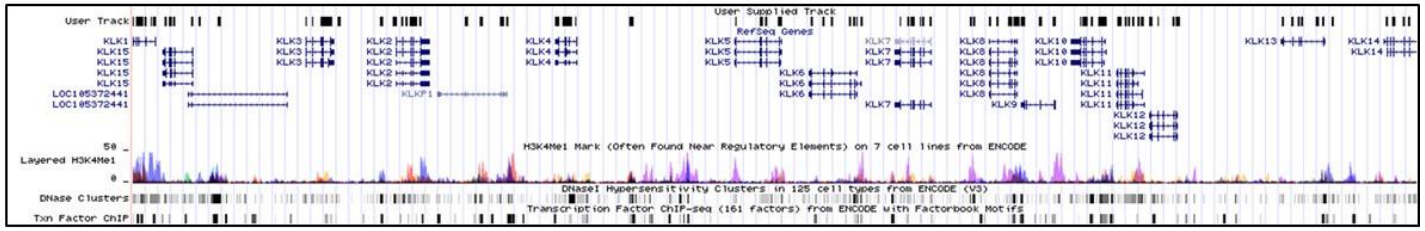


Figure S1 – Flow-chart of the strategy used to detect rare and common variants on *KLK* and *WFDC* clusters associated with male infertility. Using a DNA pooled sample approach and a high-throughput sequencing strategy, we detected in phase I 456 SNVs based on stringent filtering criteria. We then performed genotyping validation of 3 SNVs and 7 gene regions in phase II, using the same samples as in phase I. In phase III, we extended the analysis of the most promising SNVs to a further 138 controls and 95 infertility cases to allow a combined analysis of 217 controls and 238 cases.

A



B

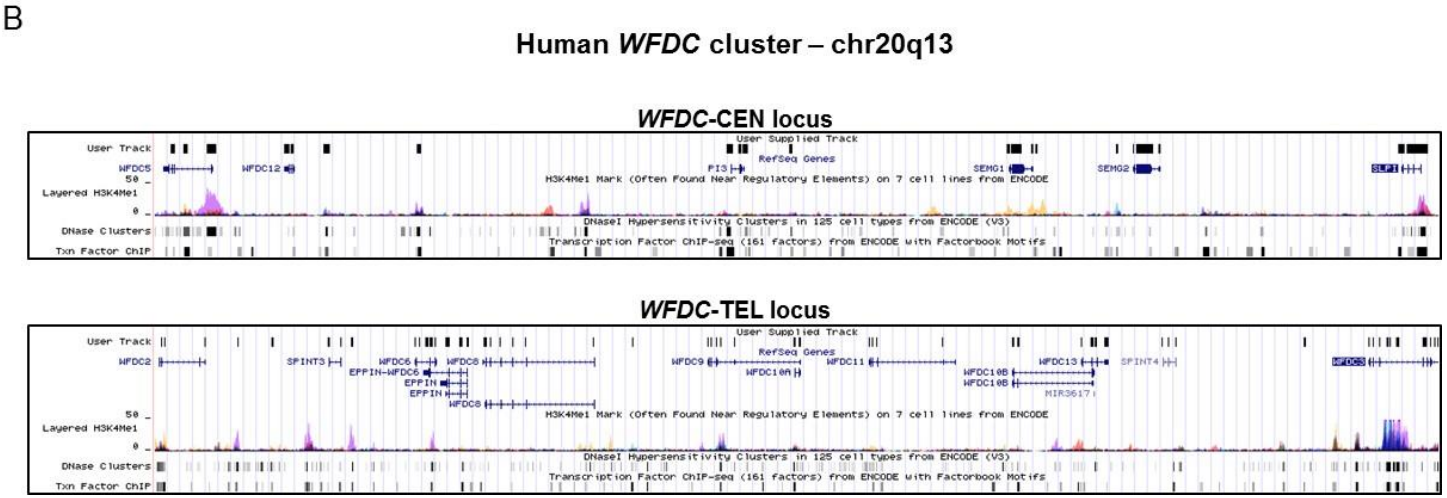


Figure S2 – Schematic representation of the human *KLK* and *WFDC* gene clusters using UCSC Genome Browser. **(A)** The human *KLK* cluster is located on chromosome 19q13.3-13.4 and includes 15 coding genes and one expressed pseudogene. **(B)** The human *WFDC* cluster located is on chromosome 20q13 and its genes are organized into two subloci (centromeric and telomeric, *WFDC*-CEN and *WFDC*-TEL, respectively), separated by 215 kb of unrelated sequence. Amplicons generated for the pilot survey, reference genes, H3K4Me1 Mark, DNase hypersensitivity and transcription factor CHIP-seq from ENCODE are shown.

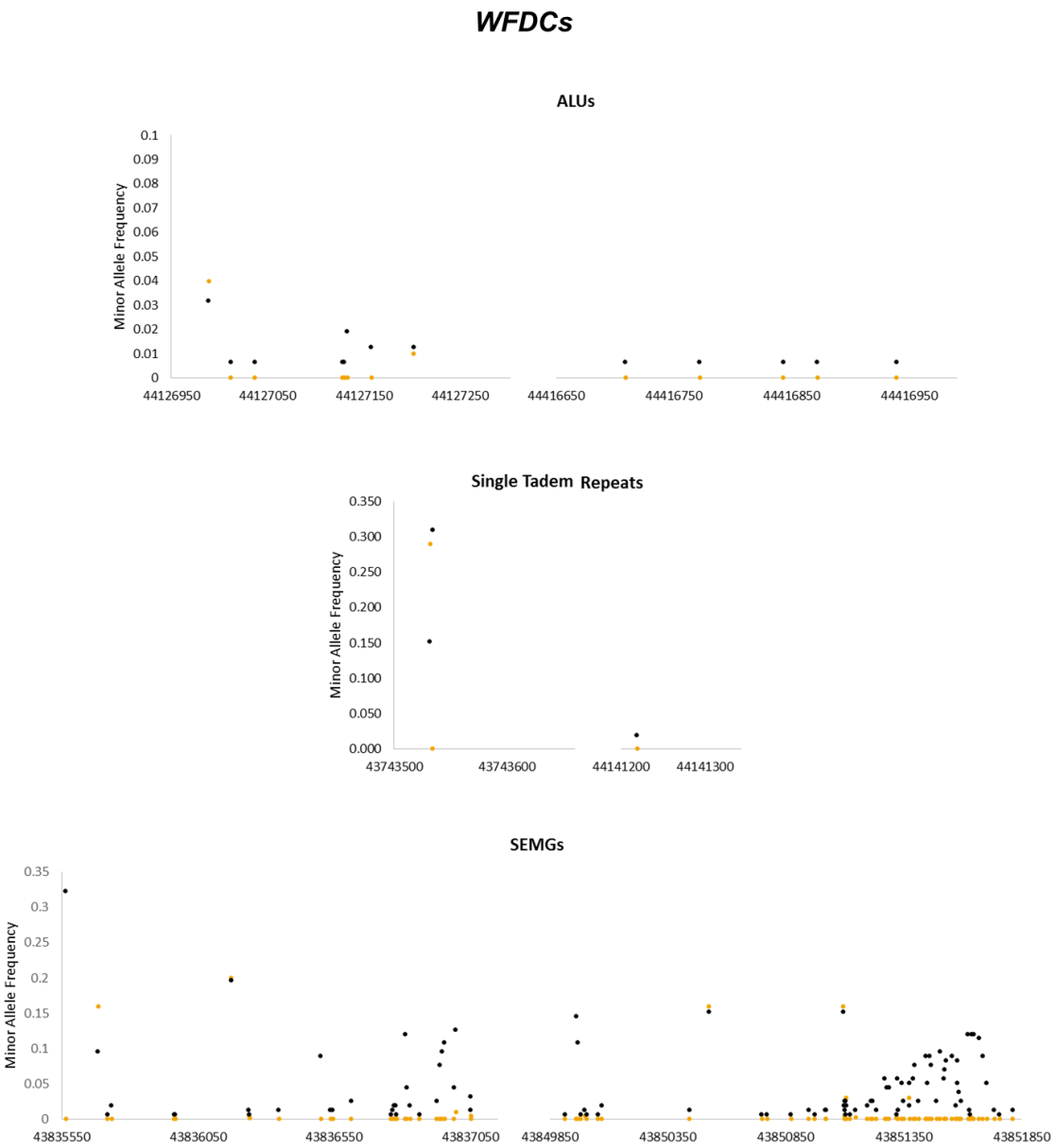


Figure S3 – *WFDCs* minor allele frequencies (MAFs) from 1000 Genomes data and control pooled sequencing in repetitive regions. Allele frequency estimates obtained in pooled sequencing for the control group (black) and the described frequencies from the combined European populations from 1000 Genomes project phase III (orange).

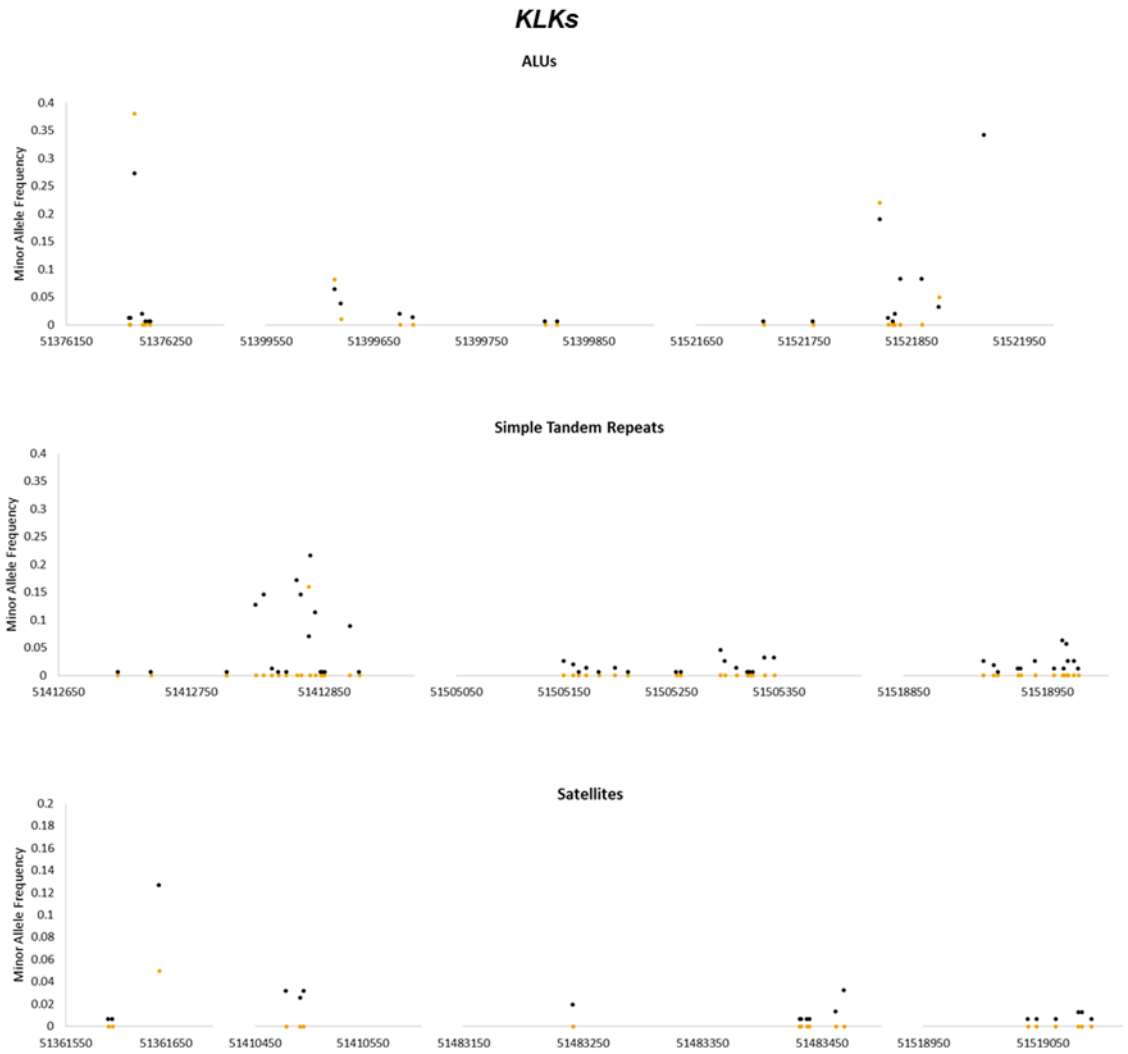


Figure S4 – *KLKs* minor allele frequencies (MAFs) from 1000 Genomes data and control pooled sequencing in repetitive regions. Allele frequency estimates obtained in pooled sequencing for the control group (black) and the described frequencies from the combined European populations from 1000 Genomes project phase III (orange).

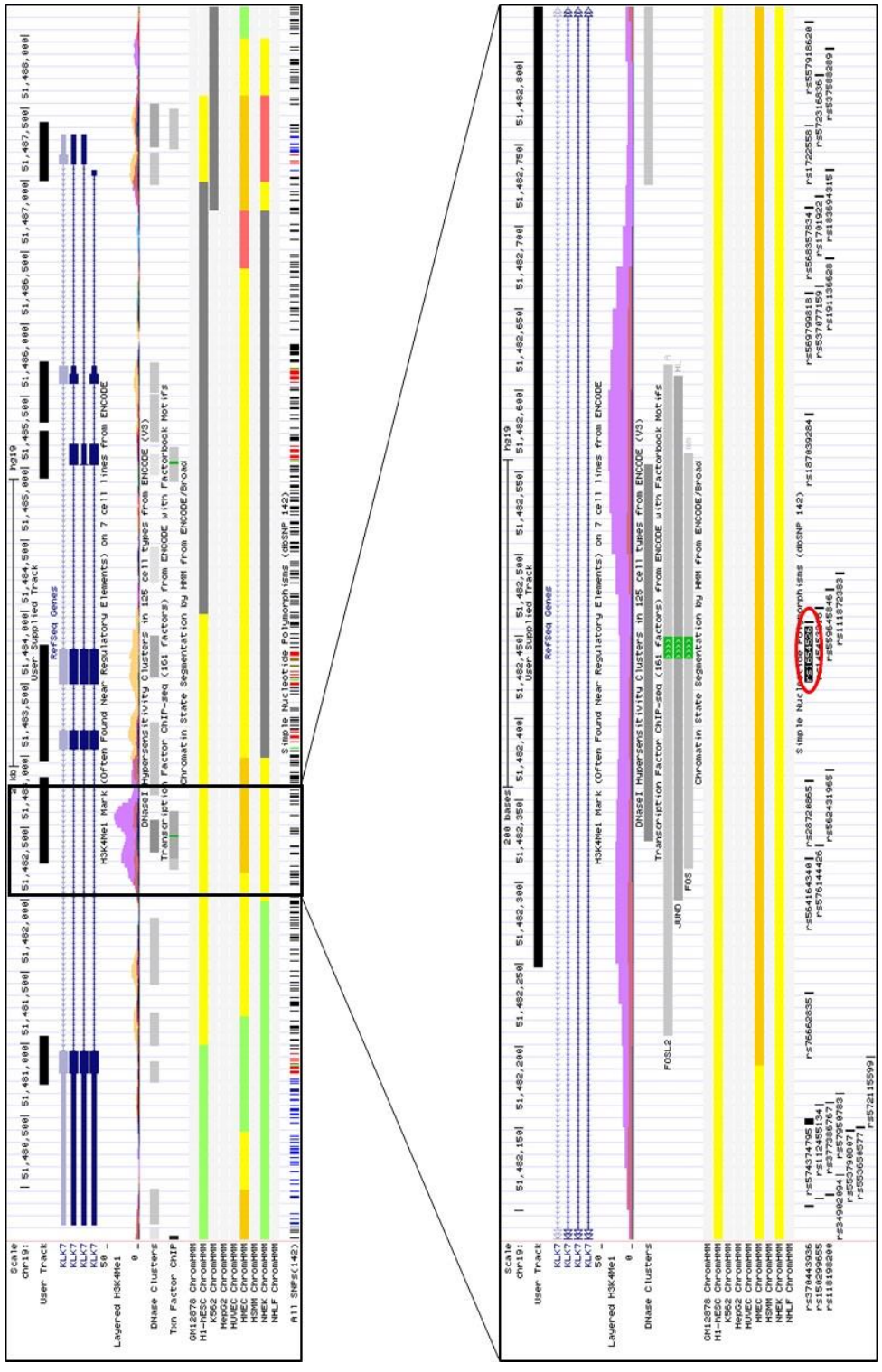


Figure S5 – Schematic representation of the human *KLK7* landscape using UCSC Genome Browser. Amplicons generated for the pilot survey, reference genes, H3K4Me1 Mark, DNase hypersensitivity, transcription factor CHIP-seq and chromatic state segmentation from ENCODE are shown in the upper image. The inset displays in detail the intron V of *KLK7* in which rs1654526 is located (highlighted by red circle).

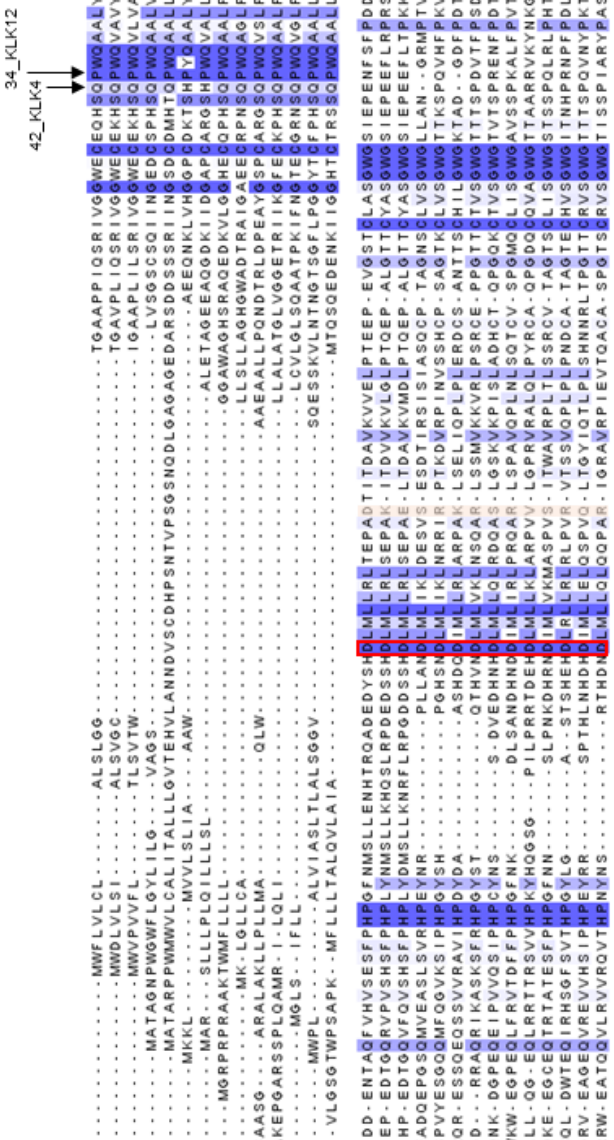


Figure S6 – Alignment of the kallikrein protein sequences. Complete conservation is shown in dark blue background, whereas partial conservation is shown on a light blue background. The catalytic residues are framed in red. Variant sites are indicated by arrows. The equivalent variants 131_KLK3 and 138_KLK14 are highlighted in pink.